Assessing the Operational Characteristics of the Individual Causal Association as a Metric of Surrogacy in the Binary Continuous Setting

Peer-reviewed author version

ONG, Fenny; MOLENBERGHS, Geert; Callegaro, Andrea; VAN DER ELST, Wim; Stijven, Florian; VERBEKE, Geert; VAN KEILEGOM, Ingrid & ALONSO ABAD, Ariel
(2024) Assessing the Operational Characteristics of the Individual Causal Association as a Metric of Surrogacy in the Binary Continuous Setting. In: Pharmaceutical statistics,.

DOI: 10.1002/pst.2437
Handle: http://hdl.handle.net/1942/44497

Received ...; | Revised ...; | Accepted ...

DOI: xxx/xxxx

MAIN PAPER

Assessing the operational characteristics of the individual causal association as a metric of surrogacy in the binary continuous setting.

Fenny Ong[1] | Geert Molenberghs[1,2] | Andrea Callegaro[3] | Wim Van der Elst[4] | Florian Stijven[2] | Geert Verbeke[2] | Ingrid Van Keilegom[5] | Ariel Alonso*[2]

[1]I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium
[2]I-BioStat, KU Leuven, Leuven, Belgium
[3]GSK Vaccines, Rixensart, Belgium
[4]The Janssen Pharmaceutical companies of Johnson & Johnson, Beerse, Belgium
[5]ORSTAT, KU Leuven, Leuven, Belgium

Correspondence
*Ariel Alonso, I-BioStat KU Leuven, B-3000 Leuven, Belgium. Email: ariel.alonsoabad@kuleuven.be

Summary

In a causal inference framework, a new metric has been proposed to quantify surrogacy for a continuous putative surrogate and a binary true endpoint, based on information theory. The proposed metric, termed the individual causal association (ICA), was quantified using a joint causal inference model for the corresponding potential outcomes. Due to the non-identifiability inherent in this type of models, a sensitivity analysis was introduced to study the behavior of the ICA as a function of the non-identifiable parameters characterizing the aforementioned model. In this scenario, to reduce uncertainty, several plausible yet untestable assumptions like monotonicity, independence, conditional independence or homogeneous variance-covariance, are often incorporated into the analysis. We assess the robustness of the methodology regarding these simplifying assumptions via simulation. The practical implications of the findings are demonstrated in the analysis of a randomized clinical trial evaluating an inactivated quadrivalent influenza vaccine.

KEYWORDS:
Causal inference, Homoscedasticity, Information theory, Monotonicity, Surrogate endpoint

1 | INTRODUCTION

Randomized clinical trials (RCTs) are the gold standard for evaluating the efficacy of new drugs and vaccines. A critical aspect of both the validity and practical feasibility of an RCT is the selection of an appropriate clinical endpoint. Often, using the most clinically relevant outcome, known as the true endpoint, requires a lengthy follow-up period and/or a large sample size, rendering the study unfeasible. In such cases, surrogate endpoints—outcomes that can be measured earlier, more conveniently, or more

frequently—may be used instead. However, before surrogate endpoints can be used as substitutes for true endpoints, they must undergo rigorous statistical evaluation. In vaccine trials, for instance, immune response measures frequently serve as surrogate endpoints for true endpoints such as protection from infection, hospitalization, intensive care admission, or mortality. These measures of immune response, known as surrogates or correlates of protection (CoP), must demonstrate accurate prediction of the vaccine's effect on the true endpoint before they can be used in RCTs for market authorization.

The assessment of surrogate endpoints has been challenging, and for over 30 years, researchers have developed statistical methods to carry out the evaluation exercise. Early efforts to develop validation methods focused on the single trial setting (STS), where information on both surrogate and true endpoints comes from a single clinical trial. However, these methods, based on expected causal treatment effects, have theoretical and practical limitations due to the lack of replication at the level of these effects in the STS.[1,2,3] To overcome this, a meta-analytic framework was introduced, solving the replication problem but requiring data from multiple trials, a resource often scarce in early drug development stages when surrogate endpoints are most crucial.[4] Consequently, developing methods for the STS has remained a priority in the field.

The evaluation of surrogate endpoints has also lacked a unified framework, leading to varied definitions and metrics. Alonso and Molenberghs[5,6] introduced an information-theoretic definition of surrogacy and the Individual Causal Association (ICA), a metric to quantify it. While the ICA is model-independent, estimating it requires a causal inference model to describe the distribution of the corresponding potential outcomes. It is worth noting that until fairly recently, such causal inference models and appropriate quantifications of the ICA had only been developed when both outcomes were either continuous or binary.[7,8] Alonso et al.[9] extended this methodology to scenarios with binary true endpoints and continuous surrogates, crucial in vaccine trials. In line with prior studies conducted in different settings,[7,8] these authors handled the non-identifiability associated with the proposed causal inference model via sensitivity analysis. As part of the sensitivity analysis, it is possible to include simplifying assumptions. Incorporating these assumptions has the potential to reduce the computational burden, to render certain aspects of the model identifiable, and to lead to more precise conclusions. However, it is essential to recognize that, in principle, these assumptions can also yield erroneous results if they are not valid. Therefore, it is crucial to examine how the methodology behaves when such misspecification occurs.

In Section 2, the ICA is introduced and in Section 3, the joint causal inference model is described. The sensitivity analysis previously proposed to tackle the identifiability issues is presented in Section 4. In Section 5, a simulation study is designed to evaluate the behavior of the proposed method and the performance of the algorithm to assess the ICA. The robustness of the methodology with respect to some widely used assumptions is studied and the main findings of the simulations are discussed in detail. The practical implications of the findings are illustrated using data from a randomized clinical trial evaluating an inactivated quadrivalent influenza vaccine in Section 6. Finally, we conclude with some remarks in Section 7.

2 | INFORMATION-THEORETIC CAUSAL-INFERENCE FRAMEWORK

2.1 | Definition of Surrogacy

In the following, the evaluation of the surrogate endpoint is conducted in the STS, i.e., using data from a single randomized clinical trial within a well-defined population. Additionally, the framework assumes the evaluation of only two treatments ($Z = \{0, 1\}$) employing a parallel study design. Consistent with the Neyman-Rubin potential outcomes model, it is presumed that each patient has four potential outcomes represented by $\boldsymbol{Y} = (T_0, T_1, S_0, S_1)^T$. Here, T_z and S_z denote the potential outcomes for the true and surrogate endpoint under treatment $z = 0$ or $z = 1$.

The implementation of the Neyman-Rubin model in this paper relies on two fundamental assumptions.[10] First, the Stable Unit Treatment Value Assumption (SUTVA)—which encompasses the absence of interference and hidden variations in treatment—establishes a link between the observed outcomes and the potential outcomes as expressed by the equation:

$$(S, T)^T = Z \cdot (S_1, T_1)^T + (1 - Z) \cdot (S_0, T_0)^T.$$

Second, the Full Exchangeability Assumption asserts that potential outcomes are independent of the assigned treatment, denoted as $(T_0, T_1, S_0, S_1)^T \perp Z$. In a randomized trial, full exchangeability is inherently guaranteed. However, SUTVA is an unverifiable assumption that requires justification through subject matter knowledge. Throughout the remainder of this paper, we make both assumptions.

The vector representing the individual causal treatment effects is defined as $\boldsymbol{\Delta} = (\Delta T, \Delta S)^T$, with $\Delta T = T_1 - T_0$ and $\Delta S = S_1 - S_0$. This formulation naturally leads to the following definition of surrogacy in the STS[6]:

Definition 1. In the STS, we shall say that S is a good surrogate for T if ΔS conveys a substantial amount of information on ΔT.

The metric that quantifies the extent of “shared" information is referred to as the individual causal association (ICA) and in the following section this concept is discussed in detail.

2.2 | Individual Causal Association

Entropy (denoted by H) and mutual information (denoted by I) are key concepts in information theory. Entropy assesses the uncertainty or information content of a random variable, while mutual information quantifies the shared information between two random variables. Consequently, mutual information serves as a suitable measure for defining the ICA in this context. Indeed, mutual information is always non-negative, attaining zero if and only if ΔT and ΔS are independent. Additionally, it is also symmetric and invariant under bijective transformations. Despite its appealing mathematical properties, the interpretability of the mutual information is hindered due to the absence of an upper bound. This limitation is addressed by mapping $I(\Delta S; \Delta T)$ into the unit interval, where zero corresponds to independence, and one implies the presence of a deterministic relation between

the individual causal treatment effects. Alonso et al.,[9] building upon the concept of multivariate dependence defined by Joe,[11] proposed the following quantification of the ICA for a binary true endpoint and a continuous surrogate:

$$R^2_H(\Delta T, \Delta S) = \frac{H(\Delta T) - H(\Delta T|\Delta S)}{H(\Delta T)} = \frac{I(\Delta S; \Delta T)}{H(\Delta T)}, \tag{1}$$

where $H(\Delta T)$ and $H(\Delta T|\Delta S)$ denote the entropy of ΔT and the conditional entropy of ΔT given ΔS, respectively. The previous definition of $R^2_H(\Delta T, \Delta S)$ satisfies the aforementioned mathematical properties, ensuring its interpretability. In essence, $R^2_H(\Delta T, \Delta S)$ represents the proportion of uncertainty about ΔT that is expected to be removed when ΔS is known.

3 | JOINT CAUSAL INFERENCE MODEL

In order to estimate (1) a joint causal inference model is needed to describe the distribution of $\boldsymbol{Y}$. The model introduced by Alonso et al.[9] is based on the following decomposition:

$$f(\boldsymbol{Y}) = f\left(\boldsymbol{Y}_T, \boldsymbol{Y}_S\right) = f\left(\boldsymbol{Y}_S \mid \boldsymbol{Y}_T\right) f\left(\boldsymbol{Y}_T\right), \tag{2}$$

where $\boldsymbol{Y}_T = (T_0, T_1)^T$ and $\boldsymbol{Y}_S = (S_0, S_1)^T$. The marginal distribution $f\left(\boldsymbol{Y}_T\right)$ is parameterized by $\boldsymbol{\pi} = (\pi_{ij})$, $i, j \in \{0, 1\}$ with $\pi_{ij} = P\left(T_0 = i, T_1 = j\right)$ and $\pi_{i.} = \sum_j \pi_{ij}$, $\pi_{.j} = \sum_i \pi_{ij}$. Due to the fundamental problem of causal inference, the association structure of the potential outcomes $\boldsymbol{Y}_T$ is unidentifiable. However, under SUTVA and the full exchangeability assumption, the marginal probabilities $\boldsymbol{\pi}_m = \left(\pi_{0.}, \pi_{1.}, \pi_{.0}, \pi_{.1}\right)^T$ can be estimated from the data. In addition to the marginal probabilities, further assumptions are needed to identify the distribution of $\boldsymbol{Y}_T$. For instance, after assuming a value for one of the cell probabilities, say π_{10} (the proportion of patients that respond to the control, but do not respond to the treatment), and using an estimate of $\boldsymbol{\pi}_m$, the distribution of $\boldsymbol{Y}_T$ can be fully identified.

The conditional density of $\boldsymbol{Y}_S \mid \boldsymbol{Y}_T$ in Equation (2) is modeled using a normal distribution $f(\boldsymbol{Y}_S \mid T_0 = i, T_1 = j) = \phi(\boldsymbol{Y}_S; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})$ where

$$\boldsymbol{\mu}_{ij} = \begin{pmatrix} \mu_0^{ij} \\ \mu_1^{ij} \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{ij} = \begin{pmatrix} \sigma_{00}^{ij} & \sigma_{01}^{ij} \\ \sigma_{01}^{ij} & \sigma_{11}^{ij} \end{pmatrix},$$

and $i, j \in \{0, 1\}$. For convenience, the conditional covariance between the surrogate potential outcomes will often be rewritten as $\sigma_{01}^{ij} = \sqrt{\sigma_{00}^{ij}\sigma_{11}^{ij}}\rho_{01}^{ij}$ where ρ_{01}^{ij} denotes the conditional correlation between S_0 and S_1. Marginally, under the previous distributional assumptions, the vector of potential outcomes for the surrogate is distributed as a mixture of normals $f\left(\boldsymbol{Y}_S\right) = \sum_{ij} \pi_{ij} \phi\left(\boldsymbol{Y}_S; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}\right)$.

TABLE 1 Conditional density of ΔS on ΔT

ΔT	$f(\Delta S \mid \Delta T)$
-1	$\phi\left(\Delta S; \mu_{*10}, \sigma_{*10}\right)$
0	$w_1\phi\left(\Delta S; \mu_{*00}, \sigma_{*00}\right) + w_2\phi\left(\Delta S; \mu_{*11}, \sigma_{*11}\right)$
1	$\phi\left(\Delta S; \mu_{*01}, \sigma_{*01}\right)$

Similarly, the joint distribution of $(\Delta T, \Delta S)$ can be written as $f\left(\Delta T, \Delta S\right) = f\left(\Delta S|\Delta T\right) f\left(\Delta T\right)$. In the previous expression, $f\left(\Delta T\right)$ is a multinomial distribution with parameters:

$$\pi_k^{\Delta T} = P\left(\Delta T = k\right) = \sum_{i,j\,:\,j-i=k} \pi_{ij} \text{ for } k \in \{-1, 0, 1\}.$$

Similar to the distribution of $\boldsymbol{Y}_T$, the distribution of ΔT is not identifiable from the data. However, once π_{10} is specified, the distribution becomes identifiable. Furthermore, $f(\Delta S|\Delta T)$ follows the set of distribution given in Table 1 , where:

$$\mu_{*ij} = \boldsymbol{a}^T \begin{pmatrix} \mu_0^{ij} \\ \mu_1^{ij} \end{pmatrix} = \mu_1^{ij} - \mu_0^{ij} = \alpha_{ij}, \tag{3}$$

$$\sigma_{*ij} = \boldsymbol{a}^T \boldsymbol{\Sigma}_{ij} \boldsymbol{a} = \sigma_{00}^{ij} + \sigma_{11}^{ij} - 2\sqrt{\sigma_{00}^{ij}\sigma_{11}^{ij}}\rho_{01}^{ij}, \tag{4}$$

$$w_1 = \frac{\pi_{00}}{\pi_{00} + \pi_{11}}, \qquad w_2 = 1 - w_1, \tag{5}$$

and $\boldsymbol{a}^T = (-1, 1)$. Marginally, the density of ΔS has the form of a mixture of normal distributions $f(\Delta S) = \sum_{ij} \pi_{ij}\phi\left(\Delta S; \mu_{*ij}, \sigma_{*ij}\right)$. Given the previous modeling assumptions, equation (1) can be computed. However, to do so, one must first estimate $f\left(\Delta T, \Delta S\right)$ using the available data.[9]

4 | IDENTIFIABILITY ISSUES

Both $\boldsymbol{Y}$ and $\boldsymbol{\Delta}$ are unobservable, rendering R_H^2 unidentifiable. To address this issue, Alonso et al.[9] proposed a sensitivity analysis based on an algorithm that assesses the ICA across a wide range of plausible distributions for $\boldsymbol{Y}$. Interested readers can refer to this publication for an in-depth discussion of the methodology and its theoretical foundation. For the sake of completeness, we provide a general outline of the procedure below. Essentially, Alonso et al.'s approach involves evaluating the ICA using the following algorithm:

1. Sample a value for π_{10} using a uniform distribution from a user predefined interval $[\pi_a, \pi_b]$ (the default is $\pi_{10} \sim \text{unif}(\pi_a = 0, \pi_b = 1)$)

2. Sample a value for the association parameter ρ_{01}^{ij} using a uniform distribution from a user predefined interval $[\rho_a, \rho_b]$ (the default is $\rho_{01}^{ij} \sim \text{unif}(\rho_a = -1, \rho_b = 1)$)
3. Determine the distribution of $\boldsymbol{Y}_T$ ($\boldsymbol{\pi} = (\pi_{ij})$) compatible with the data, using π_{10} and the estimated marginal probabilities $\boldsymbol{\pi}_m$
4. Estimate the means μ_0^{ij} and μ_1^{ij} as well as the variances σ_{00}^{ij} and σ_{11}^{ij} by fitting the models:

$$S_0 \sim \sum_{ij} \pi_{ij} \phi\left(\mu_0^{ij}, \sigma_{00}\right), \tag{6}$$

$$S_1 \sim \sum_{ij} \pi_{ij} \phi\left(\mu_1^{ij}, \sigma_{11}\right). \tag{7}$$

   to the data in the control and treatment groups, respectively, and keeping the π_{ij} parameters estimated in step 3 fixed
5. Using the transformations (3)–(5), estimate the distributions in Table 1 based on the estimates of μ_0^{ij}, μ_1^{ij}, σ_{00}^{ij}, and σ_{11}^{ij} obtained in step 4, the estimates of π_{ij} obtained in step 3 and the values of ρ_{01}^{ij} generated in step 2
6. Calculate the R_H^2
7. Repeat steps 1–6 M times

Following the proposal by Vansteelandt et al,[12] these authors identify two sources of uncertainty when assessing the ICA: (1) *imprecision* due to finite sample size and (2) *ignorance* due to unidentifiability.[9] Together, these two components constitute the overall *uncertainty* associated with the estimation of the ICA. Let $\boldsymbol{\theta}_U = \left(\pi_{10}, \rho_{01}^{00}, \rho_{01}^{01}, \rho_{01}^{10}, \rho_{01}^{11}\right)^T$ with $\boldsymbol{\theta}_U \in \Omega$ and $\Omega = [0,1] \times (-1,1)^4$ be the vector of unidentifiable parameters associated with the previously introduced causal inference model, and let $\boldsymbol{\theta}_I$ be the vector of identifiable parameters. Note that $\boldsymbol{\theta}_I$ refers to parameters that become estimable from the observed data when $\boldsymbol{\theta}_U$ is fixed to a value. To assess the *ignorance* component, we study the behavior of ICA as a function of $\boldsymbol{\theta}_U$. It is important to point out that the aforementioned algorithm only addresses the second component, and hence, it assumes that the sample size is large enough to render the first component negligible. The imprecision due to the finite sample size will be formally taken into account later in the case study analysis in Section 6.

The algorithm produces a set of values for the ICA that can be summarized using frequency distributions and intervals of ignorance. It is important to point out that, when substantive knowledge on the unidentifiable parameters is available, this could be incorporated into the algorithm by using biologically plausible intervals for $[\pi_a, \pi_b]$ and/or $[\rho_a, \rho_b]$. Alonso et al.[9] proposed to summarize the distribution of the ICA using two types intervals: the range [min(ICA), max(ICA)] and the $(1-\alpha)$ symmetric density interval (SDI). While the former offers an estimate of the maximum and minimum value of R_H^2, the latter corresponds to the central part of a distribution and it is computed based on the corresponding $\alpha/2$ and $1-\alpha/2$ quantiles. It is important to point out that, in general, the $(1-\alpha) \times 100\%$ SDI cannot be interpreted as a confidence interval as it only indicates that, for all available R_H^2 values, $(1-\alpha) \times 100\%$ of them fall within the lower and upper bound of that interval.

When implementing the previous algorithm, Alonso et al.[9] assumed that $\sigma_{00}^{ij} = \sigma_{00}$ and $\sigma_{11}^{ij} = \sigma_{11}$. This type of homoscedasticity assumption was made based on technical arguments. In fact, it can be shown that when the means and variances of a finite mixture model are different in all components, like in (6)–(7), then the model leads to an infinite likelihood. This can only be solved by keeping the variances of the components away from zero and to achieve this goal, the aforementioned homoscedasticity assumption is often made. In general, finite mixture models with equal variances allow to describe many different functional shapes and, therefore, it is plausible to assume that a four-component mixture model will be flexible enough to produce good approximations for the distribution of $S \mid Z = 0$ and $S \mid Z = 1$ in many cases.[9]

5 | SIMULATION STUDY

A simulation study was conducted to assess the behavior of the ICA introduced in Section 2 and the performance of the sensitivity analysis outlined in Section 4. The primary objective was to evaluate the impact of incorporating plausible yet untestable assumptions in the sensitivity analysis. Two key scenarios were considered in the simulation study: one where the association between the individual causal treatment effect on the surrogate and the true endpoint is small, and another where it is large.

5.1 | Simulation design

Generating data with a given R_H^2 value is challenging due to its complex dependence on the distribution of $\boldsymbol{Y}$. To emulate real-world scenarios, we chose distributions of $\boldsymbol{Y}$ derived from actual data. Our focus is on the evaluation of vaccine candidates; therefore, we based our data generation on the case study presented in Section 6. To that end, we applied the sensitivity analysis outlined in the previous section to this real case study, as detailed in Section 6. This analysis provided us with a collection of R_H^2 values consistent with the available data. We then selected two values to represent small and large associations between the individual causal treatment effect on the true and surrogate endpoints. Finally, we generated 300 datasets, each containing a total of $N = 2000$ subjects, evenly distributed across treatment groups, using the distributions associated with these ICA values. Details on how we generated the simulation data sets can be found in the Supplementary Material (A).

In general, we assumed homoscedasticity, meaning $\sigma_{00}^{ij} = \sigma_{00}$ and $\sigma_{11}^{ij} = \sigma_{11}$, and positive conditional correlations between surrogate potential outcomes, i.e., $\rho_{01}^{ij} > 0$ for all i, j. However, certain adjustments were made in various simulation settings depending on the simulation's objectives. The complete set of parameters characterizing the distributions of $\boldsymbol{Y}$ that yielded the specific ICA value in each simulation setting can be found in the Supplementary Material (B).

For fitting finite mixture models, the selection of appropriate starting values (SVs) is critical. They are usually chosen based on observed data, like the histograms of $S \mid Z = 0$ and $S \mid Z = 1$. When properly chosen, small perturbations to these SVs do not dramatically alter conclusions, as demonstrated by Alonso et al.[9] To ensure robust model fit, it is advisable to evaluate

multiple SVs by examining the maximized likelihood. While manageable with a single data set, this process becomes challenging in simulation studies involving hundreds of data sets. Therefore, in our simulation study, we used the true parameter values as SVs to minimize noise, allowing us to assess the study objectives without additional complexities from selecting initial values. For more discussion on the selection of SVs, we refer the interested reader to Alonso et al.,[9] which provides a comprehensive analysis of the topic.

Vaccine trials have unique characteristics that may not be present in other settings. For instance, in a placebo-controlled vaccine trial, one can reasonably assume that $\Delta S > 0$. To ensure that the simulation exercise remains inclusive and that the conclusions are not overly specific to a particular scenario, a series of simulation studies were also conducted using distributions based on a schizophrenia data set. The results of these simulations align with the findings of the simulation study rooted in the Influenza Vaccine trial. The detailed results of the latter are provided in the Supplementary Material (C).

5.2 | Simulation results

General analysis

Figure 1 displays the frequency distribution of the ICA, obtained after applying the sensitivity analysis as outlined in Section 4. Some important conclusions can be drawn. First, when the true value of the ICA is small ($R_H^2 = 0.2313$), the R_H^2 values from the sensitivity analysis are predominantly clustered in the first half of the unit interval. Second, when the true value of the ICA is large ($R_H^2 = 0.7869$), the R_H^2 values are spread across the second half of the unit interval, demonstrating that in this setting, the ICA can take considerably larger values. Complementing this observation, Table 2 demonstrates that the symmetric density interval (SDI) reduces the length of the intervals while maintaining the empirical coverage within a reasonable range. The length of the corresponding ignorance intervals indicates how precisely the sensitivity analysis reflects the true value of R_H^2, whereas the empirical coverage is the percentage of cases out of the 300 simulations where the true value of the ICA was contained in the calculated interval. In this simulation study, the average intervals used to summarize the results were always computed by taking the average of both the lower and upper bounds of the intervals obtained across the 300 simulated data sets. It is important to point out that the standard deviations of both the lower and upper bounds were low, providing further justification for using these average intervals to summarize the results. All the intervals given in the subsequent discussion will be average intervals. The results support the use of SDI as a probably better option to summarize the behavior of R_H^2, providing a more balanced and informative approach. Careful selection of the α value is essential to reduce the length of the interval while ensuring reasonable empirical coverage. In general, the methodology seems to be able to discriminate between a poor-performing surrogate and a promising one.

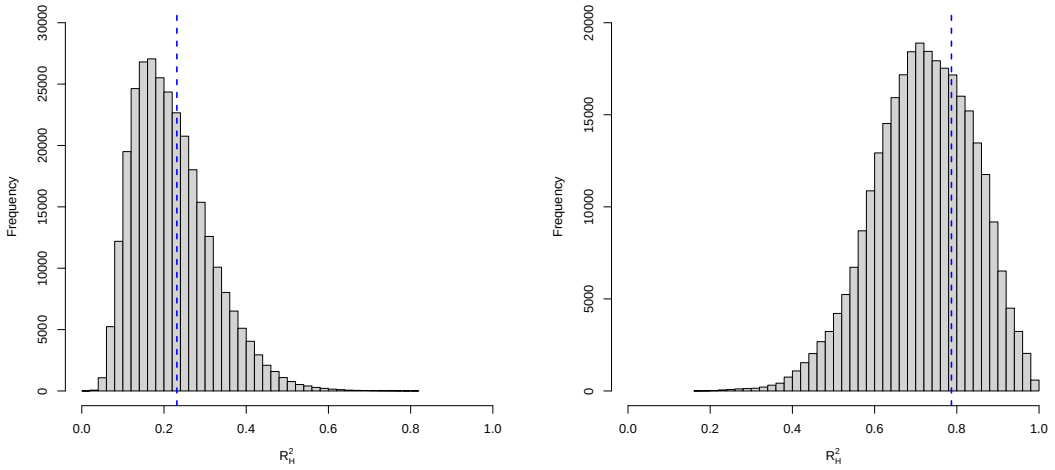


FIGURE 1 Frequency distribution of R^2_H obtained from the sensitivity analysis of 300 simulation data sets. The true values of small ICA ($R^2_H = 0.2313$) and large ICA ($R^2_H = 0.7869$) are indicated by the dashed blue vertical line.

TABLE 2 The ignorance interval and coverage of R^2_H obtained from the sensitivity analysis of 300 simulation data sets. Li: the average of lower intervals from 300 data sets; Ui: the average of upper intervals from 300 data sets. Coverage: percentage of cases (out of 300 data sets) where the true R^2_H is included in the corresponding interval.

Interval	True $R^2_H = 0.2313$		True $R^2_H = 0.7869$	
	[Li, Ui]	Coverage	[Li, Ui]	Coverage
Range	[0.07, 0.54]	100	[0.41, 0.98]	100
80% SDI	[0.12, 0.34]	99.67	[0.57, 0.87]	99
95% SDI	[0.09, 0.41]	100	[0.50, 0.92]	100

Homoscedasticity assumption

As stated in Section 4, Alonso et al.[9] proposed to fit models (6)–(7) under the assumption of homoscedasticity, i.e., assuming $\sigma^{ij}_{00} = \sigma_{00}$ and $\sigma^{ij}_{11} = \sigma_{11}$. As previously mentioned, this assumption is standard practice to prevent an infinite likelihood when fitting finite mixture Gaussian models. However, this assumption cannot be tested in practice, emphasizing the need to evaluate the methodology's robustness when this assumption is violated. To achieve this objective, we modified the simulation by generating data sets with different σ^{ij}_{00} and σ^{ij}_{11}, respectively. The generated data sets were then analyzed assuming homoscedasticity. In the subsequent discussion, coupled with findings from another simulation study using distributions based on the schizophrenia data set, it is demonstrated that the methodology is rather robust and can produce reliable results even in scenarios where the homoscedasticity assumption is not entirely satisfied.

Figure 2 shows the ICA frequency distributions obtained when homoscedasticity is wrongly assumed. In the small ICA setting ($R^2_H = 0.1589$), the range interval for the ICA is $[0.03, 0.50]$ with 100% empirical coverage, while the 80% and 95% SDI are [0.07, 0.25] and [0.05, 0.33] with an empirical coverage of 95% and 100%, respectively. It is clear that in this scenario,

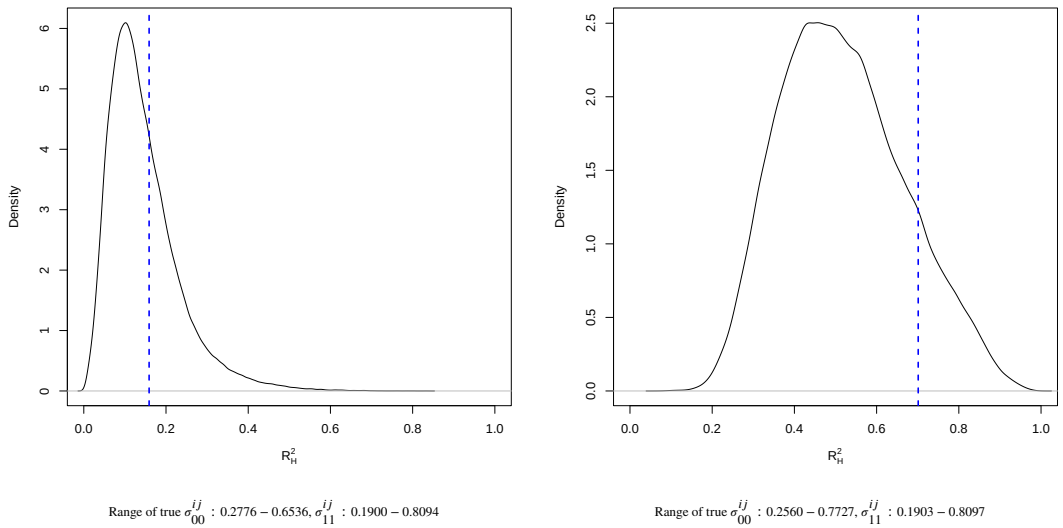


FIGURE 2 Frequency distribution of R^2_H obtained from the sensitivity analysis of 300 simulation data sets wrongly assuming homoscedasticity. The true values of small ICA ($R^2_H = 0.1589$) and large ICA ($R^2_H = 0.7012$) are indicated by the dashed blue vertical line.

the methodology manages to detect a poorly performing surrogate. In the large ICA setting ($R^2_H = 0.7012$), the range interval is [0.23, 0.94] with a 100% empirical coverage, while the 80% and 95% SDI are [0.35, 0.71] and [0.29, 0.81] with an empirical coverage of 55% and 96.33%, respectively. In this setting, the coverage of 80% SDI is rather small, but the true value of the ICA is still included in its average interval. Additionally, upon closer examination of individual results within the simulated data sets, it becomes evident that numerous upper bounds of the intervals closely approximate the true value of the ICA. However, now the methodology leads to less conclusive results as all intervals appear to be quite wide. Considering these findings collectively, it is safe to conclude that the proposed algorithm, assuming homoscedasticity, may still yield valuable results even in cases where this assumption is not correct.

Homogeneous variance-covariance assumption

In addition to the assumption of homoscedasticity, one can also assume that the association structure is homogeneous, i.e., $\rho^{ij}_{01} = \rho_{01}$ for all $i, j \in \{0, 1\}$. We call these assumptions the homogeneous variance-covariance assumption (HVCA). Under HVCA, the model for the conditional density $\boldsymbol{Y}_S \mid \boldsymbol{Y}_T$ is simplified to the set of normal distributions $\phi(\boldsymbol{Y}_S; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{00} & \sqrt{\sigma_{00}\sigma_{11}}\rho_{01} \\ \sqrt{\sigma_{00}\sigma_{11}}\rho_{01} & \sigma_{11} \end{pmatrix}.$$

Consequently, the vector of potential outcomes for the surrogate is marginally distributed as a mixture of normals $f\left(\boldsymbol{Y}_S\right) = \sum_{ij} \pi_{ij} \phi\left(\boldsymbol{Y}_S; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}\right)$, with the same variance-covariance matrix in each component of the mixture. The HVCA is often used in

multivariate procedures such as MANOVA, discriminant function analysis, and multivariate regression. The impact of wrongly assuming HVCA is explored via simulation by comparing the results obtained under this assumption with the results obtained under the general analysis. To explore the potential gains of assuming HVCA when the assumption actually holds, we generated data under both conditions, i.e., when HVCA is valid and when it is invalid. In both settings, the data were then analyzed with and without assuming HVCA. It is important to note that the data were always generated assuming homoscedasticity and, hence, the key distinction between the different settings lies in whether the association structure is homogeneous or not.

Figure 3 illustrates that the impact of assuming HVCA is minimal on the outcome of the sensitivity analysis, regardless of the validity of the assumption. For example, in the context of small ICA, erroneously assuming HVCA yielded a range interval of $[0.02, 0.99]$ and an 80% SDI of $[0.10, 0.47]$, achieving coverages of 100% and 99.67%, respectively. Conversely, when HVCA was invalid, the general analysis (without assumptions) resulted in a range interval of $[0.02, 0.80]$ and an 80% SDI of $[0.12, 0.34]$, maintaining the same coverage rates as previously mentioned.

In cases where HVCA was correctly assumed, the range interval was $[0.06, 0.80]$, and the 80% SDI equaled $[0.09, 0.47]$, with empirical coverages of 100% and 99.67%, respectively. Furthermore, when HVCA was valid, the general analysis led to a range interval of $[0.06, 0.54]$ and an 80% SDI of $[0.12, 0.33]$, achieving coverage rates of 100% and 97%, respectively.

Therefore, despite the fact that imposing HVCA in the analysis reduces the number of unidentifiable parameters, the simulations revealed that correctly assuming HVCA does not significantly improve the results. In some cases, the coverage is improved, but this comes at the cost of wider intervals. Similarly, assuming HVCA incorrectly does not seem to distort the conclusions either.

Conditional independence

Under the HVCA, one can further assume that $\rho_{01} = 0$, i.e., $S_0 \perp S_1 \mid (T_0, T_1)$. It is also important to point out that when the assumption of conditional independence is combined with other assumptions about the probability distribution of $\boldsymbol{Y}_T$, such as monotonicity, the distribution of $\boldsymbol{Y}$ becomes fully identifiable.

Four scenarios were considered in the data generation process: two for the ICA (small and large value) and two for the conditional independence (true or false). For each of these scenarios, the data were analyzed assuming conditional independence and without making any assumptions (general analysis). Figure 4 summarizes the results. For the small ICA (true $R^2_H = 0.1492$), correctly assuming conditional independence results in a range interval for the ICA of $[0.12, 0.25]$ with 82.33% empirical coverage. For the same setting, the general analysis results in a range interval for the ICA of $[0.07, 0.54]$ with 100% empirical coverage. Interestingly, correctly assuming conditional independence seems to produce a narrower range interval but with a smaller empirical coverage. Moreover, in the setting of the large ICA (true $R^2_H = 0.7204$), correctly assuming conditional

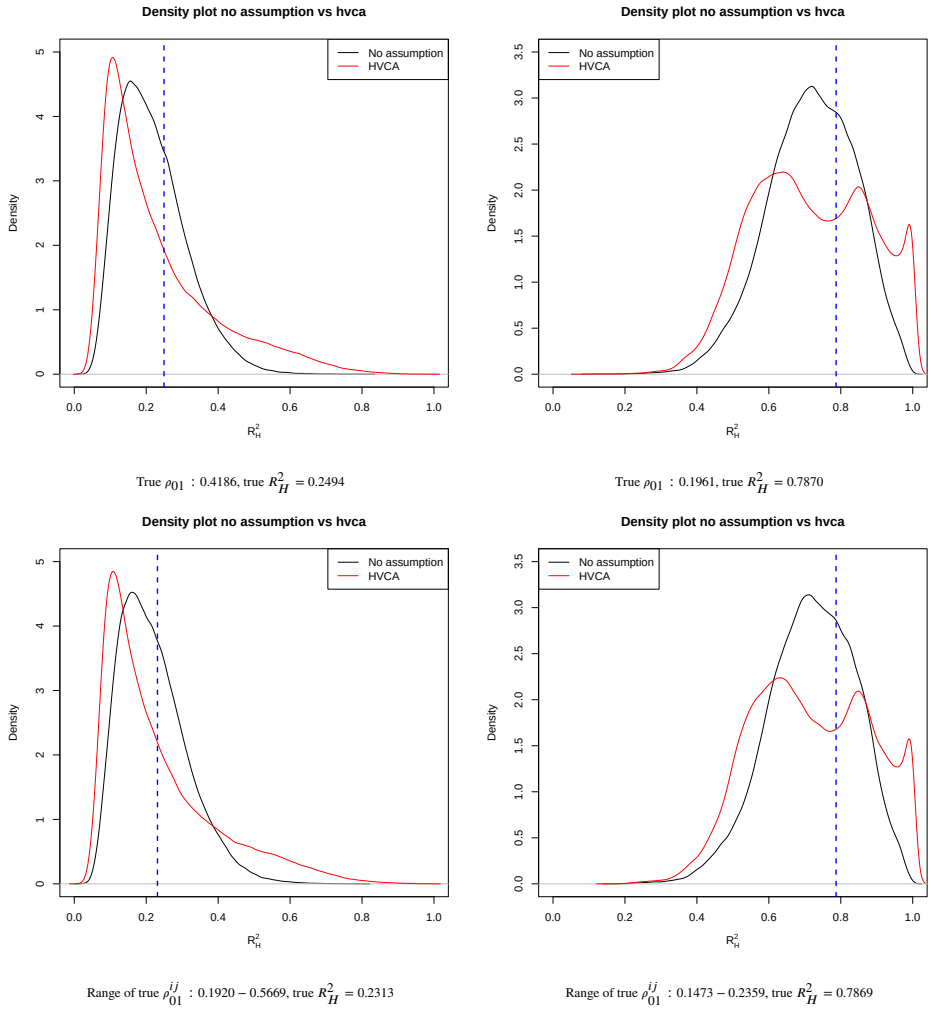


FIGURE 3 Frequency distribution of R^2_H obtained from the sensitivity analysis of 300 simulation data sets assuming HVCA when the model is correct and incorrect, respectively. The true values of the ICA are indicated by the dashed blue vertical line.

independence leads to a range interval equal to $[0.45, 0.81]$ with 99% empirical coverage, whereas the general analysis without making the assumption produces a range interval of $[0.40, 0.98]$ with an empirical coverage of 100%. In general, we may conclude that there is some value in assuming conditional independence when the assumption is correct because it may lead to substantially more precise conclusions without substantially diminishing the empirical coverage.

A different picture is obtained when conditional independence is wrongly assumed. In fact, for the small ICA (true $R^2_H = 0.2313$), when the assumption does not hold, the general analysis produces a range interval of $[0.02, 0.80]$ with an empirical coverage of 100% and the 80% SDI = $[0.12, 0.34]$ with coverage of 99.67%. However, when conditional independence is wrongly assumed, the range reduces to $[0.04, 0.50]$, but the coverage also decreases to 50.67%. Similarly, the 80% SDI = $[0.13, 0.22]$ with coverage of 33%. When the ICA is large (true $R^2_H = 0.7869$), the general analysis produces 80% SDI = $[0.57, 0.87]$ with 99% empirical coverage, whereas erroneously assuming conditional independence results in an 80% SDI = $[0.61, 0.78]$ with a

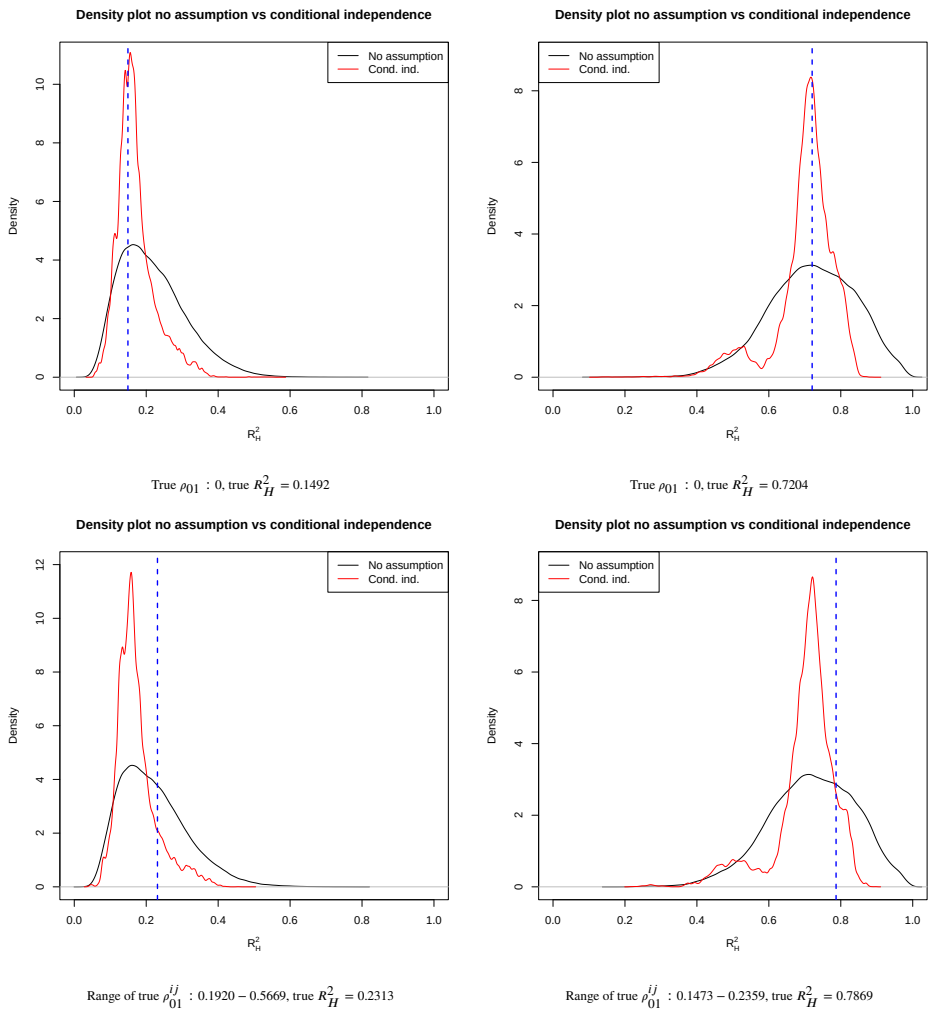


FIGURE 4 Frequency distribution of R^2_H obtained from the sensitivity analysis of 300 simulation data sets assuming conditional independence when the model is correct and incorrect, respectively. The true values of the ICA are indicated by the dashed blue vertical line.

42% empirical coverage. Clearly, wrongly assuming conditional independence may lead to narrower uncertainty intervals (minor gain for the 80% SDI), but also to substantially lower empirical coverage.

We acknowledge that conditional independence is a strong assumption. Rather than assuming that the values of $\rho^{ij}_{01} = 0$, in many applications it may be more meaningful to assume that all conditional correlations are positive or greater than some threshold, e.g., $\rho^{ij}_{01} > 0.5$. The latter assumption can be justified by considering that a baseline assessment could be used as a substitute for a counterfactual assessment under placebo. In the evaluation of changes from baseline, one measures the same outcome at different time points (baseline versus follow-up), while in the assessment of the potential outcomes, one measures the same outcome under different treatment conditions (placebo and experimental treatment). Given that high correlations are typically observed between measurements at baseline and follow-up times in clinical trials, it is reasonable to assume that the correlation between measurements of surrogate outcomes under placebo and treatment would also be high. It is also easy to show

that higher correlations between measurements imply smaller standard deviations in the change from baseline, and similarly, in the difference between the surrogate potential outcomes, indicating more precise measurement in the difference between the surrogate potential outcomes. Nonetheless, given the untestable nature of this assumption and the serious impact it may have on the results, we strongly advise always using it in the context of a sensitivity analysis and never in isolation.

Monotonicity assumption

The parameter π_{10} is the proportion of individuals in the population who would remain uninfected under a placebo but would get infected if they would receive the vaccine. In many vaccine clinical trials, the so-called assumption of monotonicity, i.e., $\pi_{10} = P(T_1 < T_0) = 0$ is often made. Indeed, in the context of a vaccine trial, under monotonicity the new vaccine cannot lead to a worse outcome than the control. Even though monotonicity seems plausible in a placebo-controlled vaccine trial, there are several possible mechanisms through which a vaccine can produce an enhanced susceptibility to virus infection or to aberrant viral pathogenesis for infections from members of different virus families.[13,14] The studies indicated that in some situations, the monotonicity assumption might have to be called into question. Furthermore, one may also argue that the well-known placebo effect could also lead to violations of monotonicity. Therefore, it is useful to study the impact of making this assumption on the evaluation exercise when it is actually false.

To assess the impact of making the monotonicity assumption on the results of the sensitivity analysis, we modified the simulation by generating data sets with increasing values of π_{10} and then analyzed them under the assumption of monotonicity. This can be done by fixing the value of π_{10} in step 1 in the proposed algorithm as 0, instead of sampling a value from a uniform distribution. For comparison, we also analyzed the datasets without imposing any assumptions.

Figure 5 provides the ICA frequency distributions obtained when the true ICA value was large. It is clear that when π_{10} is very small such that the monotonicity condition approximately holds, the analysis conducted under the assumption of monotonicity yielded results similar to the general analysis without any assumptions. However, when the value of π_{10} gets larger, the results of the sensitivity analysis assuming monotonicity become more misleading. For example, assuming monotonicity with $\pi_{10} = 0.11$ results in a range interval for the ICA of $[0.13, 0.67]$, with an empirical coverage of only 17%. Meanwhile, the range interval for the ICA in the analysis without the monotonicity assumption is $[0.16, 0.96]$ with an empirical coverage of 100%. Furthermore, when monotonicity was not assumed, the 80% and 95% SDI are $[0.50, 0.86]$ and $[0.26, 0.90]$, respectively, with both intervals having a 100% empirical coverage. When monotonicity was assumed, the 80% and 95% SDI are $[0.16, 0.40]$ and $[0.14, 0.49]$ with an empirical coverage of only 3.33% and 5%, respectively. Clearly, the previous findings indicate that the results obtained by mistakenly assuming monotonicity will likely discourage the use of a potentially valuable surrogate in this context. Similar conclusions can be drawn from the setting when $\pi_{10} = 0.16$ where all empirical coverages are zero when monotonicity is

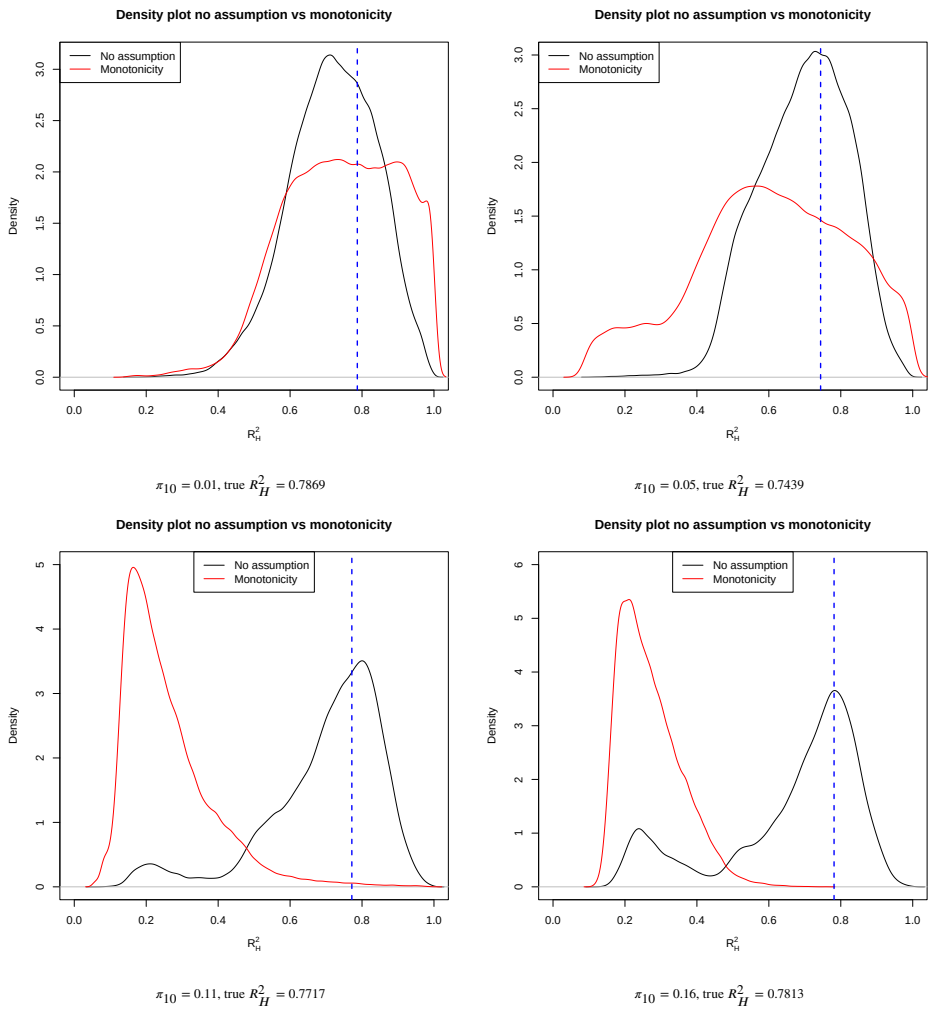


FIGURE 5 Frequency distribution of R^2_H obtained from the sensitivity analysis of 300 simulation data sets assuming monotonicity when the model is not correct. The true values of the ICA are indicated by the dashed blue vertical line.

wrongly assumed. Interestingly, when the true ICA was small, the impact of wrongly assuming monotonicity was much less pronounced (see Supplementary Material (B) for a comprehensive discussion).

The previous findings illustrate that, under minor deviations from monotonicity, the methodology still yields meaningful results when this assumption is made. However, significant deviations may have a substantial impact on the outcome of the sensitivity analysis if monotonicity is actually assumed and the true ICA value is large. It is important to point out that similar results have been observed in the binary-binary setting, i.e., when both the surrogate and true endpoints are binary random variables.[15] In general, cautious interpretation is essential, particularly when the analysis under monotonicity leads to a rather different conclusion compared to the general analysis. It is important to note that, although the assumption of monotonicity cannot be tested directly, the algorithm readily reveals that the maximum possible value of π_{10} is constrained by the minimum of $\{\pi_{1\cdot}, \pi_{\cdot 0}\}$, which can be identified from the data. As this minimum value approaches zero, the decision to assume monotonicity

in the analysis becomes more justifiable. Additionally, π_{10} is further bounded below by $\max(0, \pi_{1.} - \pi_{.1})$ and $\max(0, \pi_{.0} - \pi_{0.})$. In instances where any of these limits surpasses zero, it is possible to exclude the monotonicity assumption.

Monotonicity and conditional independence

As previously mentioned, the simultaneous application of monotonicity and conditional independence assumptions ensures the complete identifiability of the distribution of $\boldsymbol{Y}$ and, consequently, the ICA. If these assumptions are correct, then the proposed algorithm will produce the maximum likelihood estimate of the ICA. It is worth reiterating that the algorithm introduced in Section 4 does not account for the *imprecision* component when evaluating the ICA, i.e., the *imprecision* arising from the sampling variability in the parameter estimates from models (6)–(7). Meyvisch et al.[16] proposed a solution to this issue by generating bootstrap samples of the data and applying the sensitivity analysis to each sample. In this context, assuming the validity of monotonicity and conditional independence, the resulting $(1-\alpha)\times 100\%$ SDI from these bootstrap samples transforms into the traditional $(1-\alpha)\times 100\%$ percentile bootstrap confidence interval for the ICA. We generated 300 bootstrap samples for each simulated data set to implement this procedure. However, it is essential to note that employing this method in a simulation study can be computationally demanding since the sensitivity analysis must be performed for each bootstrap sample of each generated data set.

We examined the impact of applying these identifiability conditions incorrectly through simulations. We explored four distinct scenarios: when both monotonicity and conditional independence were valid ($\pi_{10} = 0.01$ and $\rho_{01} = 0$), when either monotonicity or conditional independence was valid ($\pi_{10} = 0.01$ and range $\rho_{01}^{ij} = 0.1473 - 0.2359$ or $\pi_{10} = 0.11$ and $\rho_{01} = 0$), and when both assumptions were violated ($\pi_{10} = 0.11$ and range $\rho_{01}^{ij} = 0.1473 - 0.2359$). To evaluate the methodology's performance, we calculated the relative bias of the estimated ICA and the actual coverage probability of the $(1-\alpha)\times 100\%$ bootstrap confidence intervals provided by the algorithm. Table 3 provides a summary of the findings. As anticipated, when both assumptions are valid, the average relative bias is zero, and the coverage probabilities slightly exceed the expected theoretical values. However, when one of the assumptions is invalid, the methodology yields misleading results. The data indicate that violating the monotonicity assumption results in significantly higher relative bias and lower coverage probabilities compared to violating the conditional independence assumption.

Identifiability conditions are crucial in surrogacy evaluation studies as they enable the estimation of essential surrogacy metrics. Our findings serve as a cautionary reminder about the potential consequences of relying on such assumptions when they may not be valid.

TABLE 3 Relative bias and coverage of R_H^2 obtained from the sensitivity analysis of 300 simulation data sets assuming monotonicity and conditional independence (CI).

Assumption		Relative bias of R_H^2				Coverage (%)		N
Monotonicity	CI	Min	Max	Mean	Median	80% SDI	95% SDI	
True	True	-0.8558	0.1377	0.0023	0.0264	90.11	98.17	273
True	False	-0.7010	0.0578	-0.0832	-0.0676	43.99	75.56	266
False	True	-0.9086	-0.1822	-0.7157	-0.7401	0	0	224
False	False	-0.9270	-0.2238	-0.7336	-0.7544	0	0	231

Relative bias of R_H^2: $(\hat{R}_H^2 - R_H^2)/R_H^2$, where $\hat{R}_H^2$ is the estimated R_H^2 obtained from the sensitivity analysis when monotonicity and conditional independence are assumed;
Coverage: percentage of cases where the true R_H^2 is included in the corresponding SDI;
N: total number of data sets (out of 300 simulation data sets) with available SDI. Note that not all simulation data sets are compatible with the monotonicity assumption.

6 | CASE STUDY

The data was collected from a phase-3, observer-blind, randomized, multi-country, and controlled clinical study to evaluate the efficacy of an inactivated quadrivalent influenza vaccine.[17] In our analysis, we considered a group of 1379 children whose hemagglutination inhibition (HI) antibody titer was measured 28 days after the last vaccination dose. The true endpoint was the absence of infection at the end of the study. The scientific question of interest is whether the HI antibody titer could be used as a valid CoP for the binary true endpoint. In this article, only the main results are presented and interested readers are referred to the Supplementary Material (D) for a more detailed description of how we implemented the method using our own developed R package *Surrogate*, freely available at http://cran.r-project.org/web/packages/Surrogate/.

A primary analysis of these data along with the selection procedure of the SV used to fit models (6)–(7) has been done in a previous study.[9] In the current work, we extended the analysis by considering several plausible assumptions that are frequently made in this context. To deal with the issue of imprecision due to finite sample size, we followed the approach proposed by Meyvisch et al.,[16] namely by taking bootstrap samples of the data and applying the sensitivity analysis to each of them. We generated 300 bootstrap samples in our analysis and combined all values of R_H^2 obtained from each sample using a frequency distribution. The results are summarized in Figure 6 .

The range interval, which represents the smallest and largest ICA that aligns with the available data after considering the sampling variability, spans from 0.0139 to 0.9999. The somewhat inconclusive findings obtained from the observed range are not completely unexpected. However, despite occasional small ICA values, 80% of the values surpass 0.6150. Actually, both the 80% SDI = $(0.4853, 0.8470)$ and 95% SDI = $(0.2154, 0.9229)$ provide some level of quantitative support for the use of HI antibody titer as a CoP for the absence of infection.

When the sensitivity analysis does not consider sampling variability, the resulting range interval is [0.2572, 0.9744], predictably narrower than the previous result. Additionally, the 80% SDI $(0.6299, 0.8586)$ and 95% SDI $(0.5594, 0.9254)$ reduce

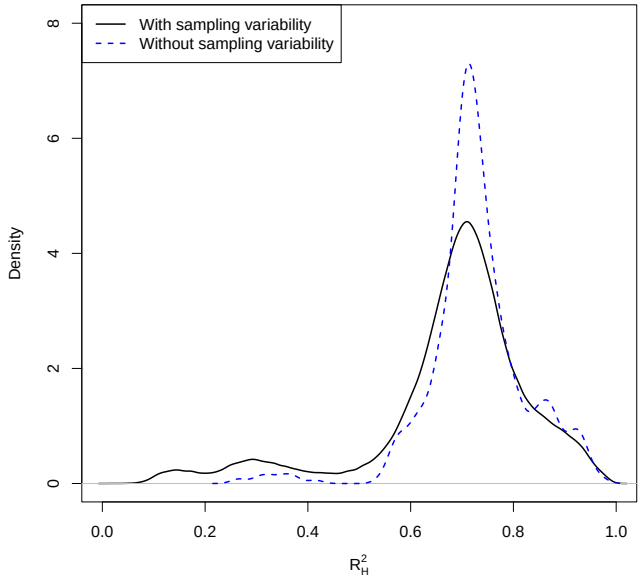


FIGURE 6 Frequency distribution of R_2^H from the case study data set with and without accounting for the sampling variability.

the length of the intervals and offer greater confidence in employing HI antibody titers as a substitute for true endpoint protection. However, as it can be seen in Figure 6 , the inclusion of sampling variability produces a more frequent occurrence of lower values for the ICA. In general, we advise always to account for sampling variability when assessing a CoP and hence, in the following, we will only discuss the results obtained using the bootstrap approach.

Despite employing a non-influenza control vaccine in the study instead of a placebo, the estimated value derived from the data for $\min\{\pi_{1\cdot}, \pi_{\cdot 0}\}$, which is an upper bound for π_{10}, was 0.0357. This indicates that the monotonicity assumption may be, at least approximately, appropriate in this case study. In addition, as the results of the simulations clearly demonstrated, assuming monotonicity when π_{10} is close to zero does not seem to lead to misleading results. As can be learned from Table 4 , assuming monotonicity leads to narrower uncertainty intervals as compared with the general analysis and under this assumption, we also found more support for the use of the CoP.

Besides monotonicity, one can make another assumption about the association structure of the potential outcomes of the true endpoint, namely the independence assumption ($T_0 \perp T_1$). Independence implies that the protection given by the control vaccine would convey no information about the protection provided by the experimental vaccine and can be formally written as $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$. Making this assumption also leads to narrower uncertainty intervals; however, one might be somewhat reluctant to justify this assumption from a biological perspective.

TABLE 4 Uncertainty intervals of R_H^2 from the sensitivity analysis of 300 bootstrap samples of the case study data set when different simplifying assumptions are incorporated in the analysis.

Assumption	R_H^2					
	Range		80% SDI		95% SDI	
	Min	Max	Min	Max	Min	Max
No assumption	0.0139	0.9999	0.4853	0.8470	0.2154	0.9229
Monotonicity	0.0729	0.9976	0.5653	0.7957	0.4759	0.9227
Independence	0.0748	0.9632	0.5588	0.7830	0.2890	0.8351
HVCA	0.0223	0.9998	0.5416	0.8207	0.2647	0.9111
Conditional independence (CI)	0.0077	0.9985	0.5536	0.8064	0.2981	0.8919
Monotonicity and CI	0.1315	0.9407	0.5793	0.7890	0.4876	0.9120
Independence and CI	0.1367	0.9049	0.5655	0.7708	0.2956	0.8273

Comparable to the findings in the simulation study, we observed that integrating the HVCA yielded results akin to those obtained from the general analysis (no assumptions). However, when conditional independence was also added, the uncertainty intervals became a bit narrower.

As mentioned earlier, when the assumption of conditional independence is coupled with monotonicity or independence, the distribution of $\boldsymbol{Y}$ and consequently, the ICA, becomes identifiable. In this case, the $(1-\alpha)\times 100\%$ SDI can be interpreted as the $(1-\alpha)\times 100\%$ bootstrap confidence interval for this parameter. For instance, assuming the simultaneous validity of monotonicity and conditional independence, the 95% SDI indicates that knowing the treatment effect on the CoP will reduce our uncertainty about the treatment effect on the true endpoint within a range of 49% to 91%. However, the results of the simulations clearly demonstrate that making these assumptions may lead to wrong conclusions when they are violated. Therefore, we strongly advise the use of these identifiability conditions only in the context of a sensitivity analysis and never as a unique inferential tool. In this example, assuming monotonicity and conditional independence produces a narrower range interval than assuming monotonicity alone. Interestingly, adding the conditional independence assumption to the monotonicity assumption seems to have a little impact on the 80% and 95% SDI uncertainty intervals.

We believe that the results obtained under the monotonicity assumption are likely the most meaningful in this case, considering its biological plausibility and the range of π_{10} values that are compatible with the data. When the results are combined with more substantive knowledge about the characteristics of the vaccine and the role of HI antibody titers in developing protection, we may justify the use of HI antibody titers as a CoP in some situations. Ultimately, experts and health authorities must weigh the benefit/risk ratio associated with a CoP before deciding on its use in the evaluation process of a vaccine candidate.

7 | CONCLUSION

The present work intends to further study the behavior of the new metric of surrogacy proposed in the binary-continuous setting developed based on the information theory. The proposed metric, the individual causal association (ICA), quantifies the association between the individual causal treatment effects on the surrogate and true endpoint. To assess the ICA, a joint causal inference model for the potential outcomes of both variables was developed. To address the non-identifiability inherent in this type of model, Alonso et al.[9] introduced a sensitivity analysis. In this framework, simplifying assumptions can be made to reduce the number of non-identifiable parameters and to obtain more precise conclusions. In general, the methodology seems to work well when the homoscedasticity assumption is violated.

A distinct perspective surfaced when evaluating the influence of simplifying assumptions. The results obtained in the simulations and the ones presented in Table 4 clearly show that making some simplifying assumptions may have a substantial impact on the outcome of the sensitivity analysis. The key takeaway here is the necessity for caution. While certain simplifying assumptions such as homoscedasticity or the HVCA appear to exert only minor influence, others like monotonicity or conditional independence can yield more precise results when they are correct, but the reduced interval length may be misleading, as it will likely be offset by bias, when these assumptions are not valid. Given that these assumptions are inherently untestable, our simulation results strongly advocate for prioritizing the use of a sensitivity analysis as the primary inferential tool in this context.

Recently, there has been recognition that multiple immune responses, including cell-mediated immunity, may interact to provide protection against infection. Lesko and Atkinson[18] also highlighted that most biomarkers are unlikely to fully capture treatment effects alone and future trials should consider a combination of biomarkers to predict treatment effect on the clinical outcome. Given this, from a statistical perspective, there is a growing interest in developing methods and surrogacy metrics capable of incorporating multiple surrogates to predict treatment effect on the true outcome of interest. This concept has already been realized in the continuous-continuous setting.[19] Future research may focus on developing the ICA metric in the binary-continuous setting, where multiple surrogates are available to predict the true endpoint of interest.

Exploring the extension of this methodology to other types of outcomes, such as time-to-event data, would be practically valuable and it will be the subject of future research. In fact, scenarios where both surrogate and true endpoints are time-to-event outcomes are common and highly relevant, not only for evaluating vaccine candidates but also for other therapeutic areas such as oncology and cardiovascular diseases.

Another interesting avenue for future research involves comparing different methodologies through simulations. However, it is important to recognize the inherent challenges in such comparisons, given the distinct goals and merits of each methodology, as highlighted by Joffe and Greene[20] and Elliott, Li, and Taylor.[21] Despite these challenges, some studies have investigated the relative performance of different methodologies in specific scenarios. For example, Alonso et al.[22] provided valuable insights

into the connection between the information-theoretic causal inference approach and the meta-analytic framework when both outcomes are continuous. Additionally, when both outcomes are binary, Alonso et al.[8] demonstrated, using theoretical arguments, that the ICA not only possesses useful mathematical properties and an appealing interpretation but also offers a more coherent assessment of surrogacy compared to other established metrics, such as the associative and dissociative proportions defined within the principal stratification framework.[23]

FUNDING AND ACKNOWLEDGMENTS

Fenny Ong gratefully acknowledges the support from the Special Research Fund (BOF) of Hasselt University (BOF-number: BOF2OCPO3) and GlaxoSmithKline Biologicals for this study. Florian Stijven gratefully acknowledges funding from Agentschap Innoveren & Ondernemen and Janssen through a Baekeland Mandate (grant number: HBC.2022.0145).
The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

Conflict of interest

Andrea Callegaro is an employee of and holds shares in the GSK group of companies.

Author contributions

AA, GM, FO, and AC contributed to the study design and methodology. FO and WVE contributed to the development of the statistical programming and analysis of the data. FO, AA, and GM prepared the first draft of the manuscript. FS, GV, and IVK contributed to the revision of the manuscript. All authors approved the final version.

Data availability statement

Anonymized individual participant data and study documents can be requested for further research from www.clinicalstudydatarequest.com.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website. An R package implementing the proposed methods is available at http://cran.r-project.org/web/packages/Surrogate/.

References

1. Prentice R. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med* 1989; 8: 431-440.

2. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; 54: 1014-1029.

3. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; 1(1): 49-67.

4. Alonso A, Bigirumurame T, Burzykowski T, et al. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Boca Raton, FL: Chapman & Hall/CRC . 2017.

5. Alonso A, Molenberghs G. Surrogate marker evaluation from an information theory perspective. *Biometrics* 2007; 63: 180-186.

6. Alonso Abad A. An information-theoretic approach for the evaluation of surrogate endpoints. *Wiley StatsRef: Statistics Reference Online* 2018: 1-7. doi: 10.1002/9781118445112.stat08157

7. Van Der Elst W, Molenberghs G, Alonso A. Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints. *Stat Med* 2016; 35: 1281-1298.

8. Alonso A, Van Der Elst W, Molenberghs G, Buyse M, Burzykowski T. An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. *Biometrics* 2016; 72: 669-677.

9. Alonso Abad A, Ong F, Stijven F, et al. An information-theoretic approach for the assessment of a continuous outcome as a surrogate for a binary true endpoint based on causal inference: Application to vaccine evaluation. *Stat Med* 2024: 1-20. doi: 10.1002/sim.9997

10. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1): 41-55.

11. Joe H. Relative entropy measures of multivariate dependence. *J Am Stat Assoc* 1989; 84(405): 157-164.

12. Vansteelandt S, Goetghebeur E, Kenward M, Molenberghs G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat Sin* 2006; 16: 953-979.

13. Huisman W, Martina B, Rimmelzwaan G, Gruters R, Osterhaus A. Vaccine-induced enhancement of viral infections. *Vaccine* 2009; 27: 505-512.

14. Huang Y, Follmann D, Nason M, et al. Effect of rAd5-vector HIV-1 preventive vaccines on HIV-1 acquisition: A participant-level meta-analysis of randomized trials. *PLoS ONE* 2015; 10(9): 1-19.

15. Alonso A, Van Der Elst W, Molenberghs G. A maximum entropy approach for the evaluation of surrogate endpoints based on causal inference. *Stat Med* 2018; 37(29): 4525-4538.

16. Meyvisch P, Alonso A, Van Der Elst W, Molenberghs G. Assessing the predictive value of a binary surrogate for a binary true endpoint based on the minimum probability of a prediction error. *Pharm Stat* 2018; 18(3): 304-315.

17. Claeys C, Zaman K, Dbaibo G, et al. Prevention of vaccine-matched and mismatched influenza in children aged 6-35 months: A multinational randomised trial across five influenza seasons. *Lancet Child Adolesc Health* 2018; 2: 338-349.

18. Lesko L, Atkinson A. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. *Annu Rev Pharmacol Toxicol* 2001; 41: 347-366.

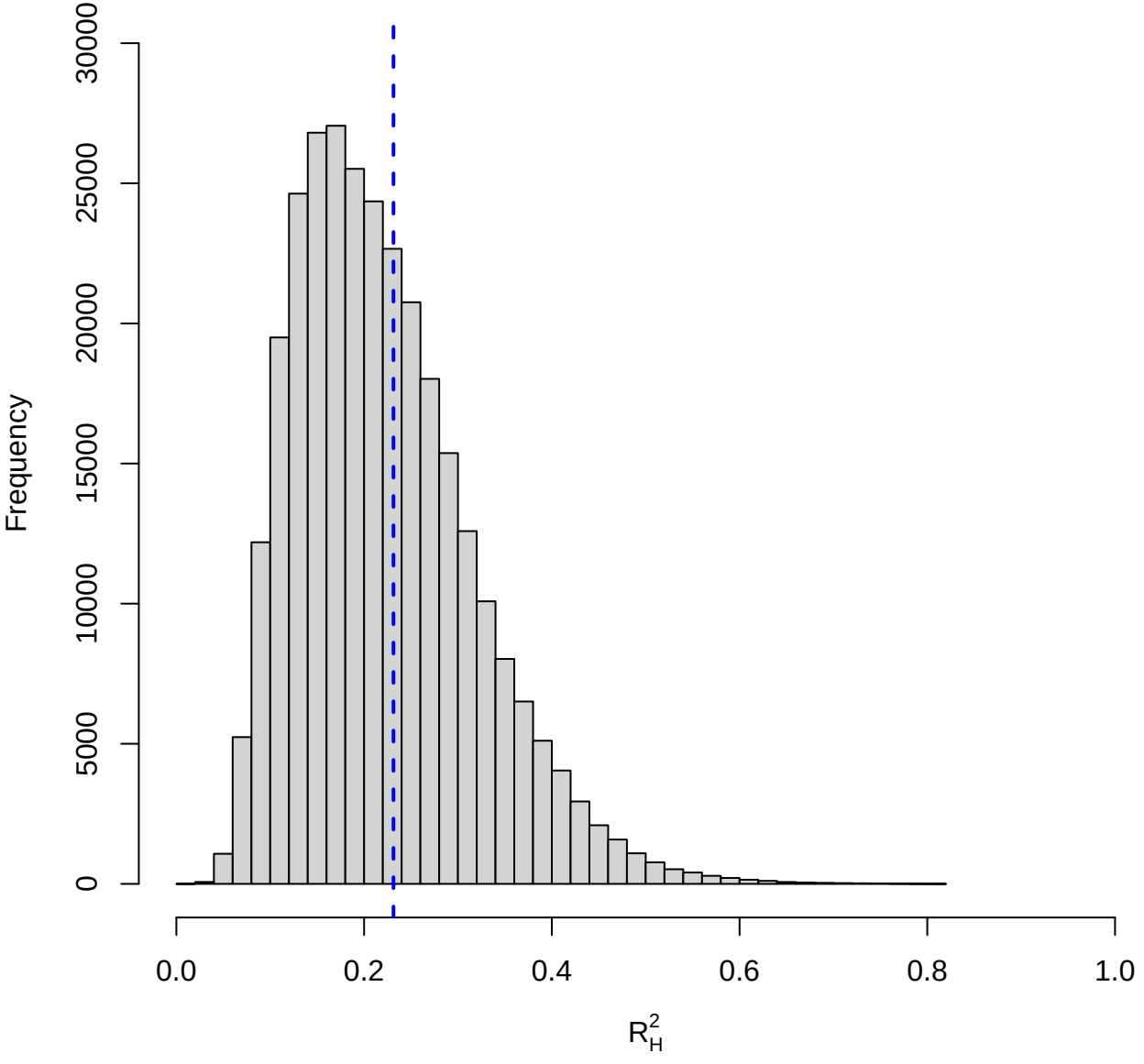
19. Van Der Elst W, Alonso Abad A, Geys H, et al. Univariate versus multivariate surrogates in the single-trial setting. *Stat Biopharm Res* 2019; 11(3): 301-310.

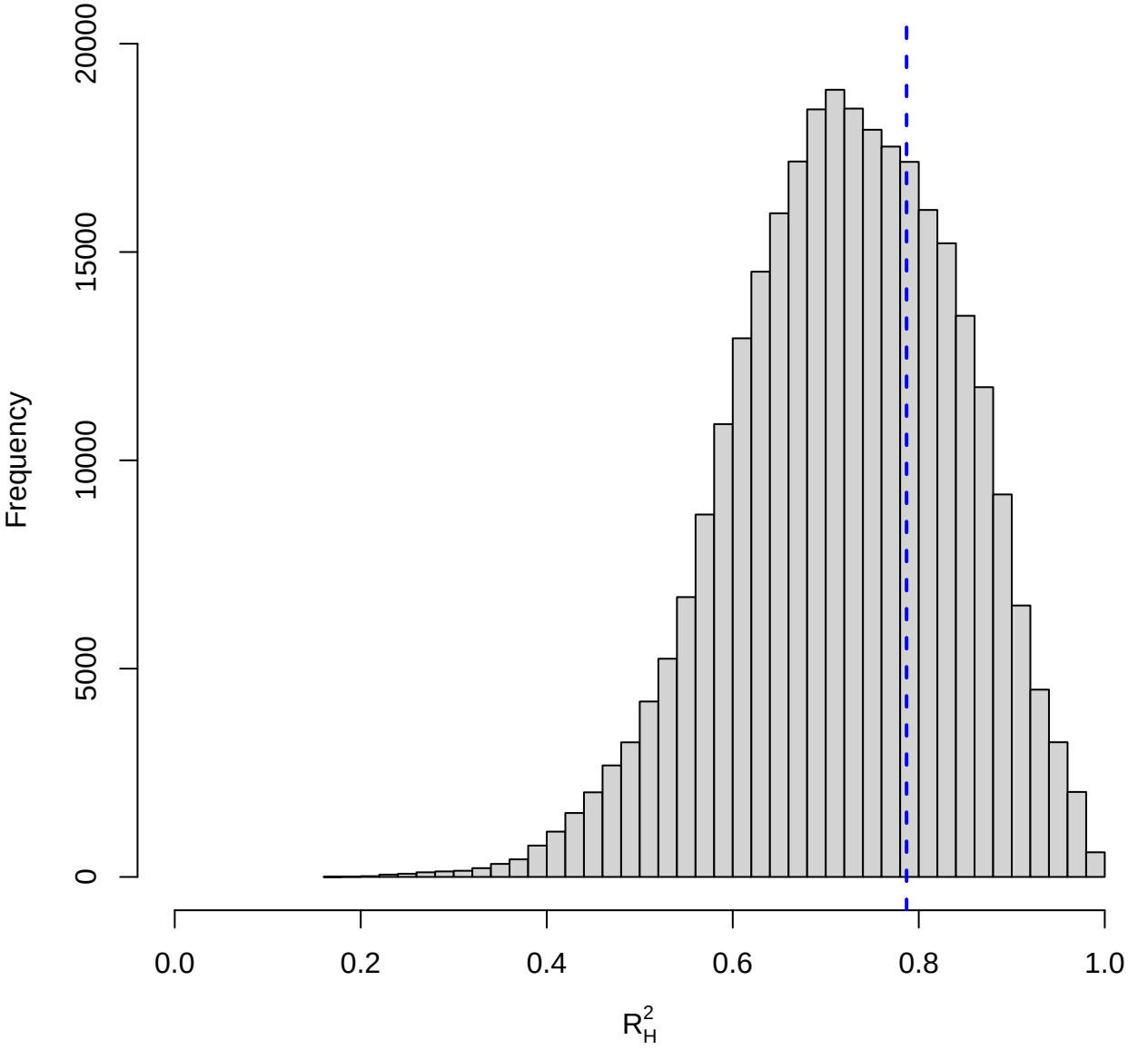
20. Joffe M, Greene T. Related causal frameworks for surrogate outcomes. *Biometrics* 2009; 65: 530-538.

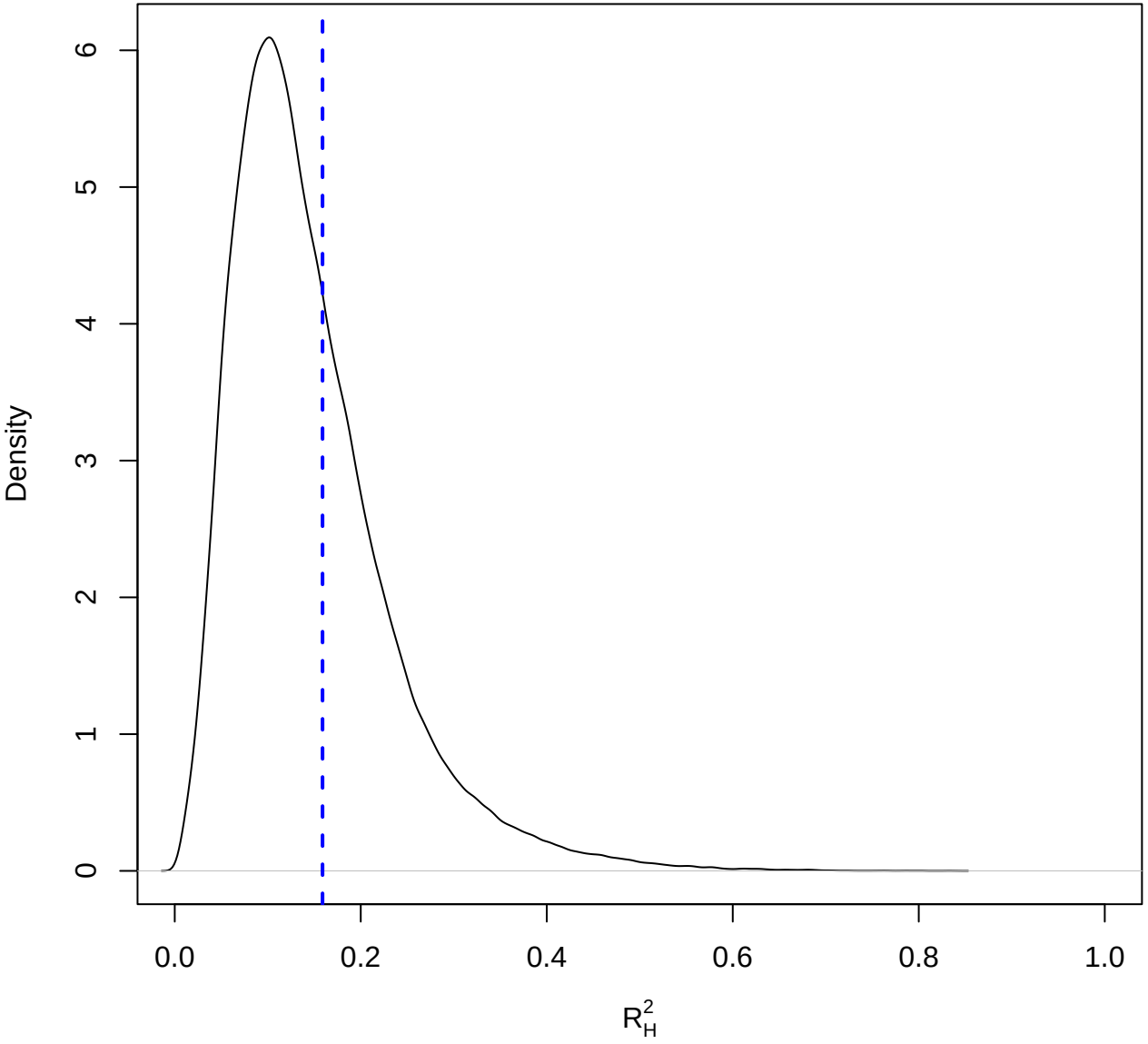
21. Elliott M, Li Y, Taylor J. Accommodating missingness when assessing surrogacy via principal stratification. *Clin Trials* 2013; 10(3): 363–377.

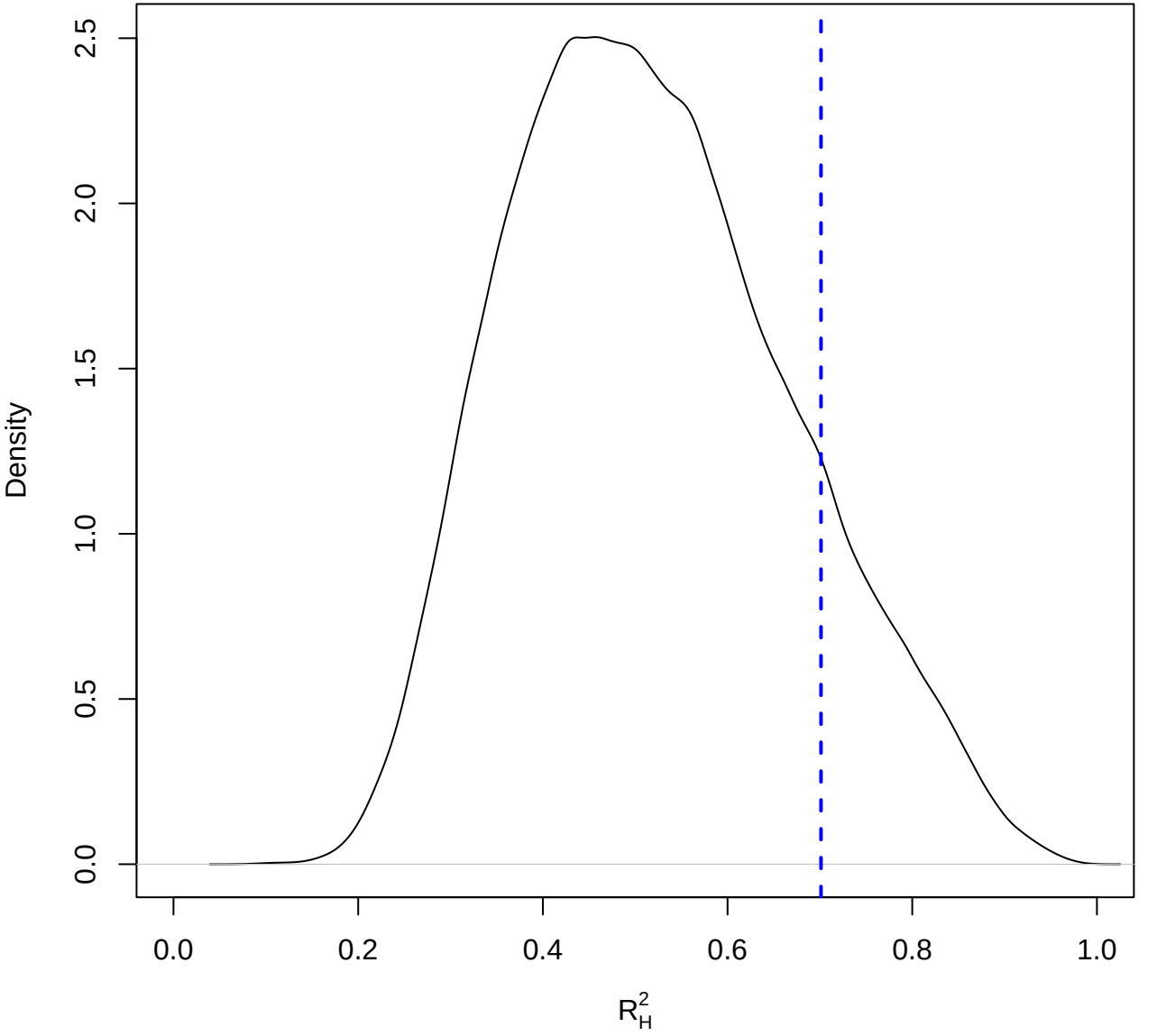
22. Alonso A, Elst V. dW, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics* 2015; 71: 15–24.

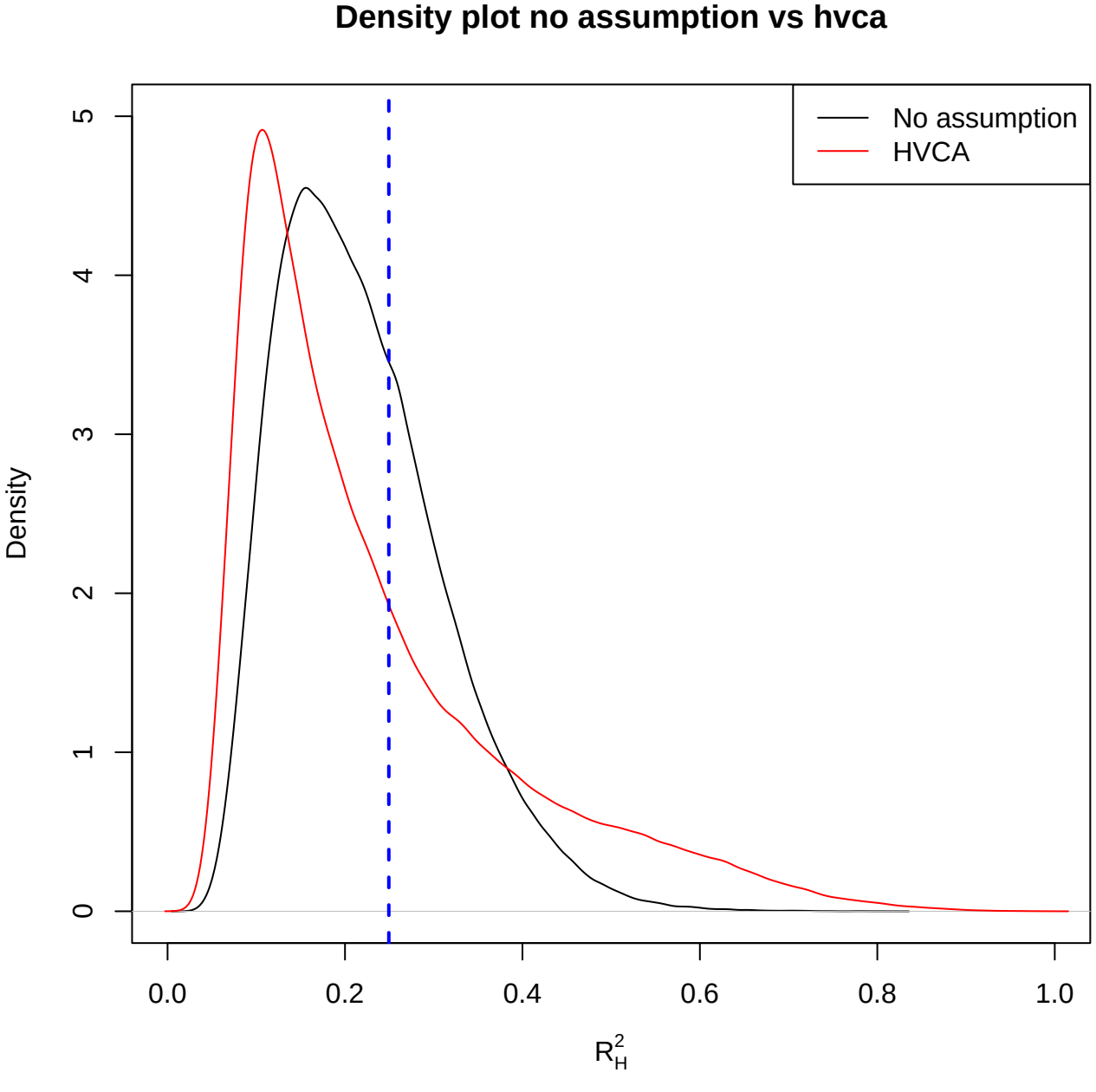
23. Frangakis C, Rubin D. Principal stratification in causal inference. *Biometrics* 2002; 58: 21-29.

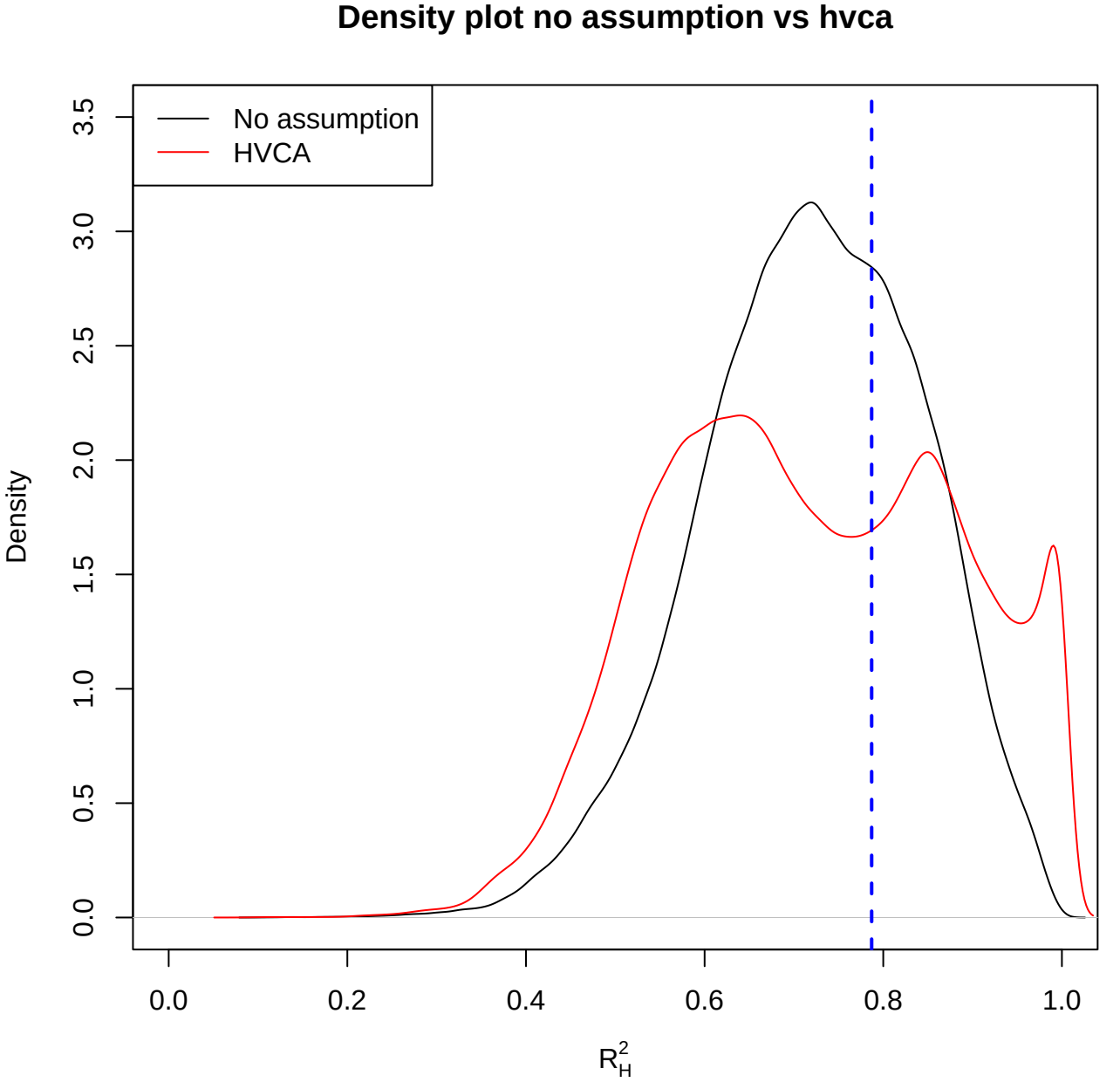

Frequency
0
5000
10000
15000
20000
25000
30000
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H

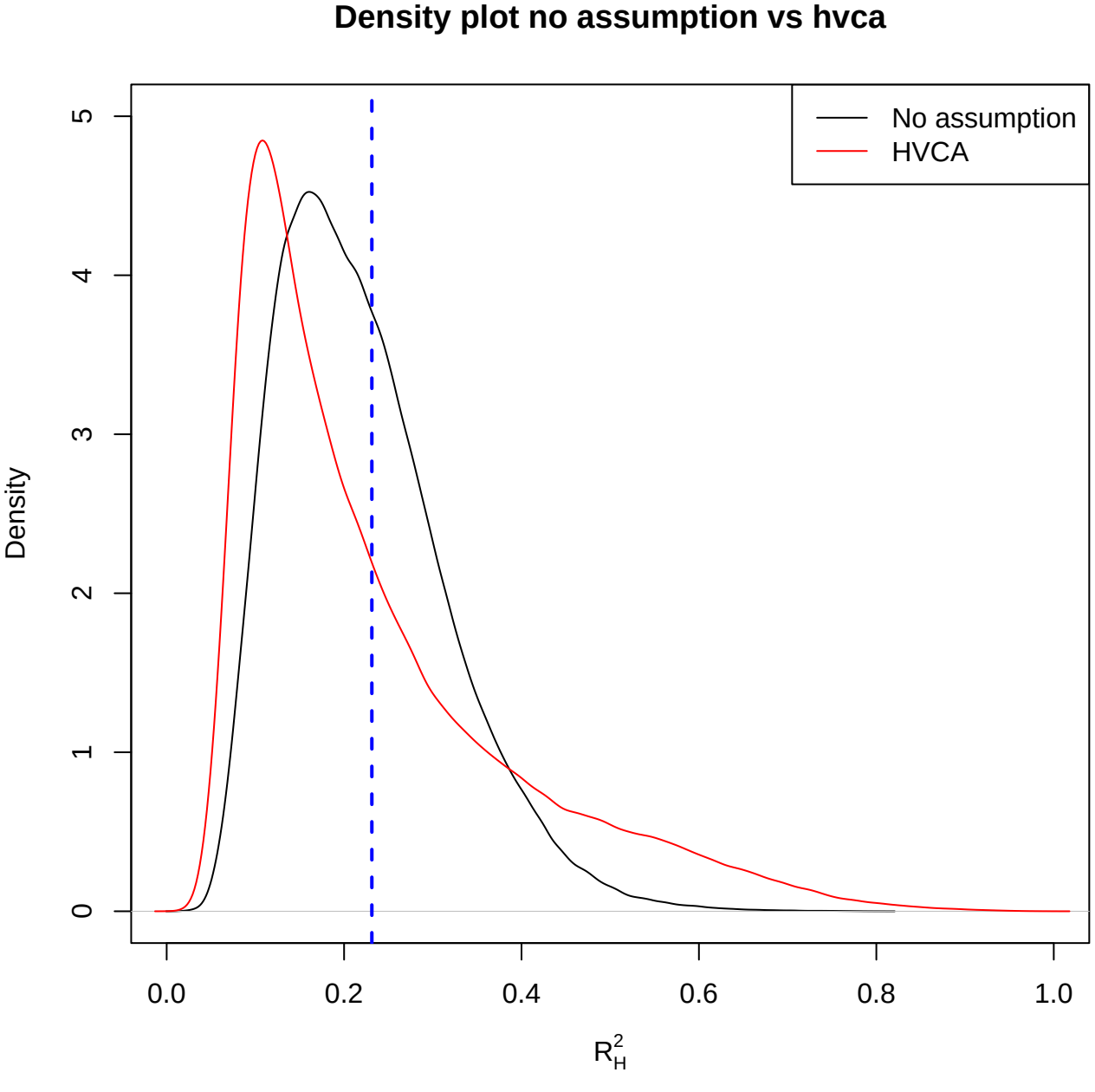

Frequency
0
5000
10000
15000
20000
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H

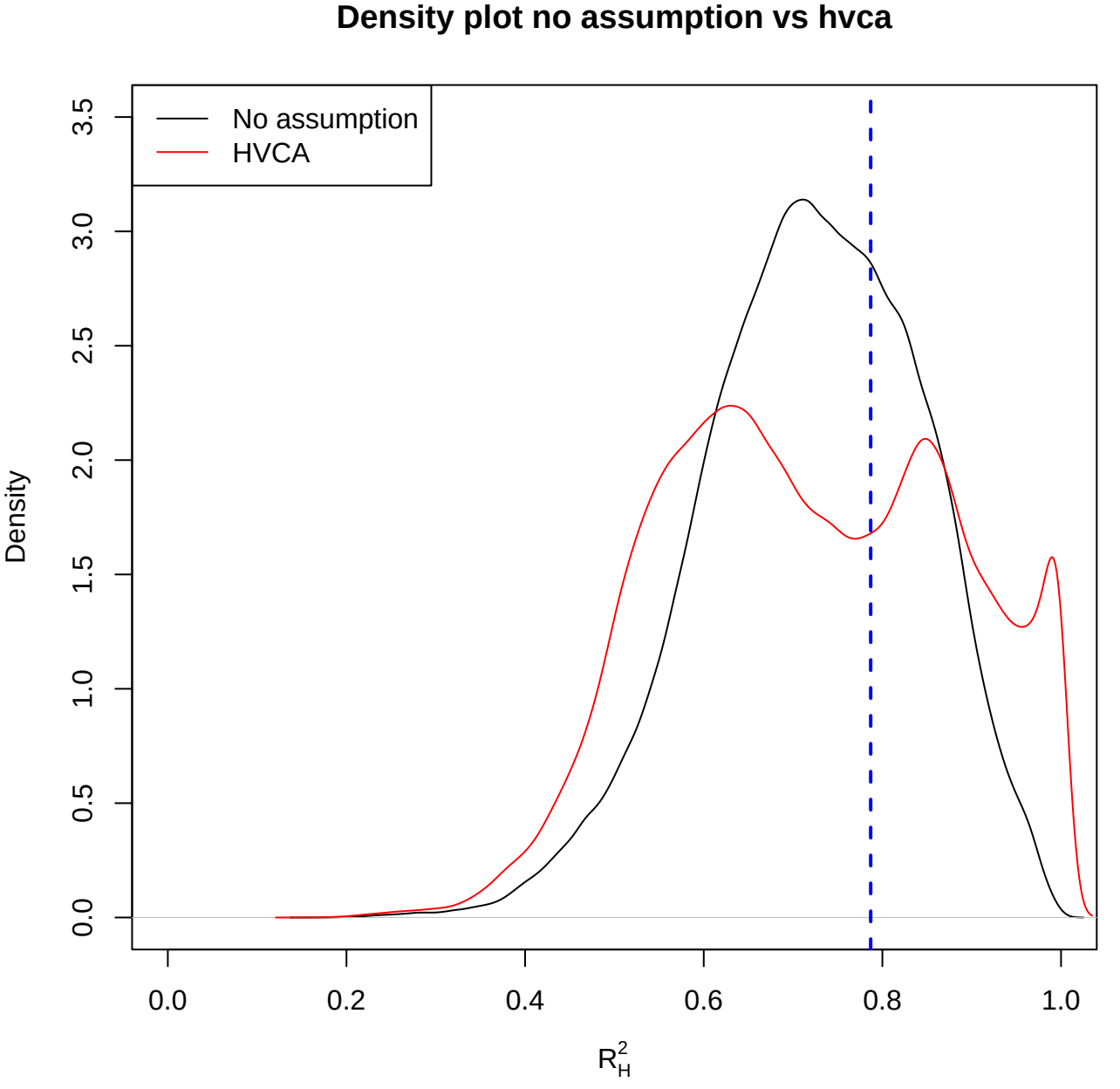

Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H


Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H

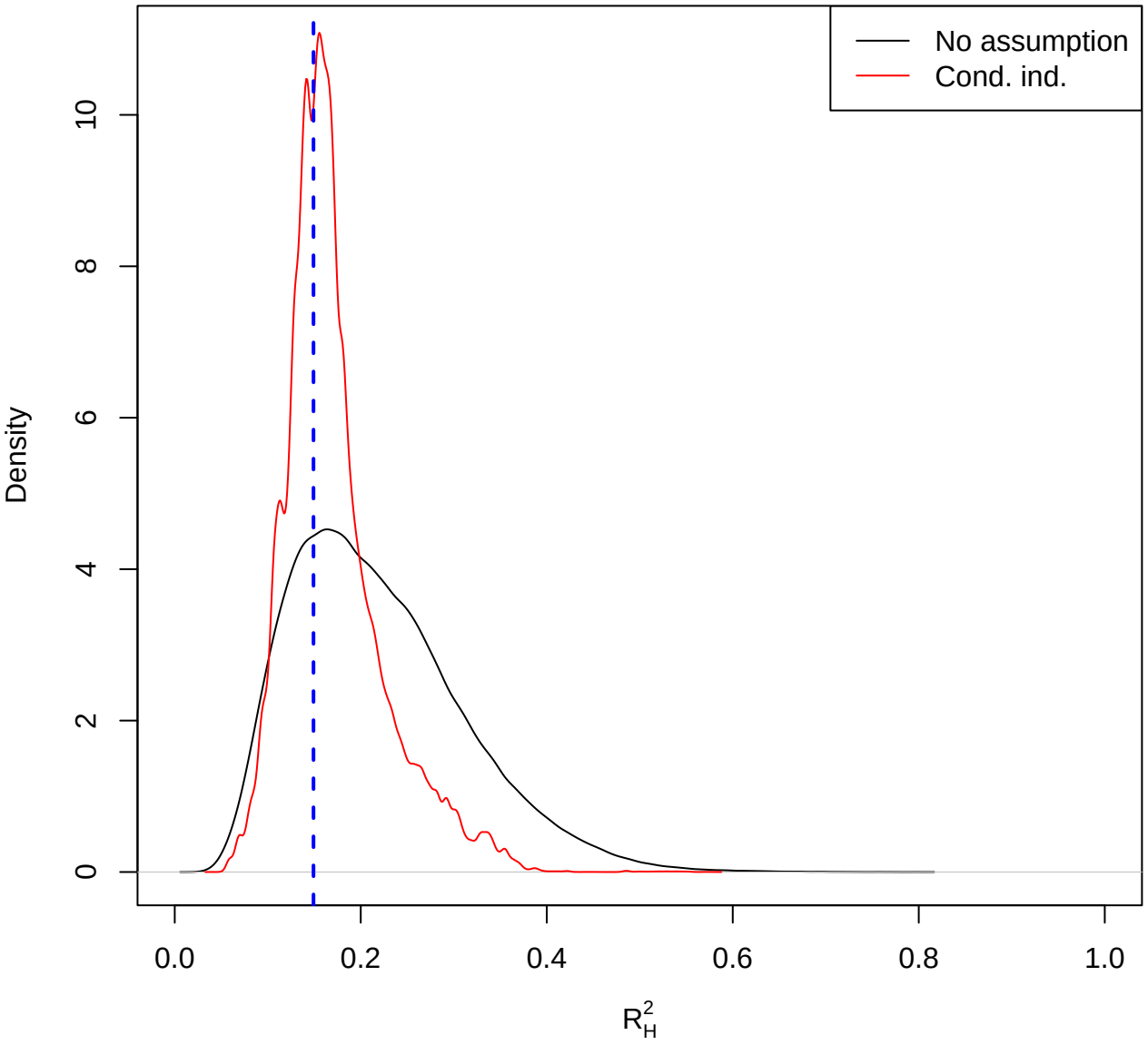

Density plot no assumption vs hvca
No assumption
HVCA
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H


Density plot no assumption vs hvca
No assumption
HVCA
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H

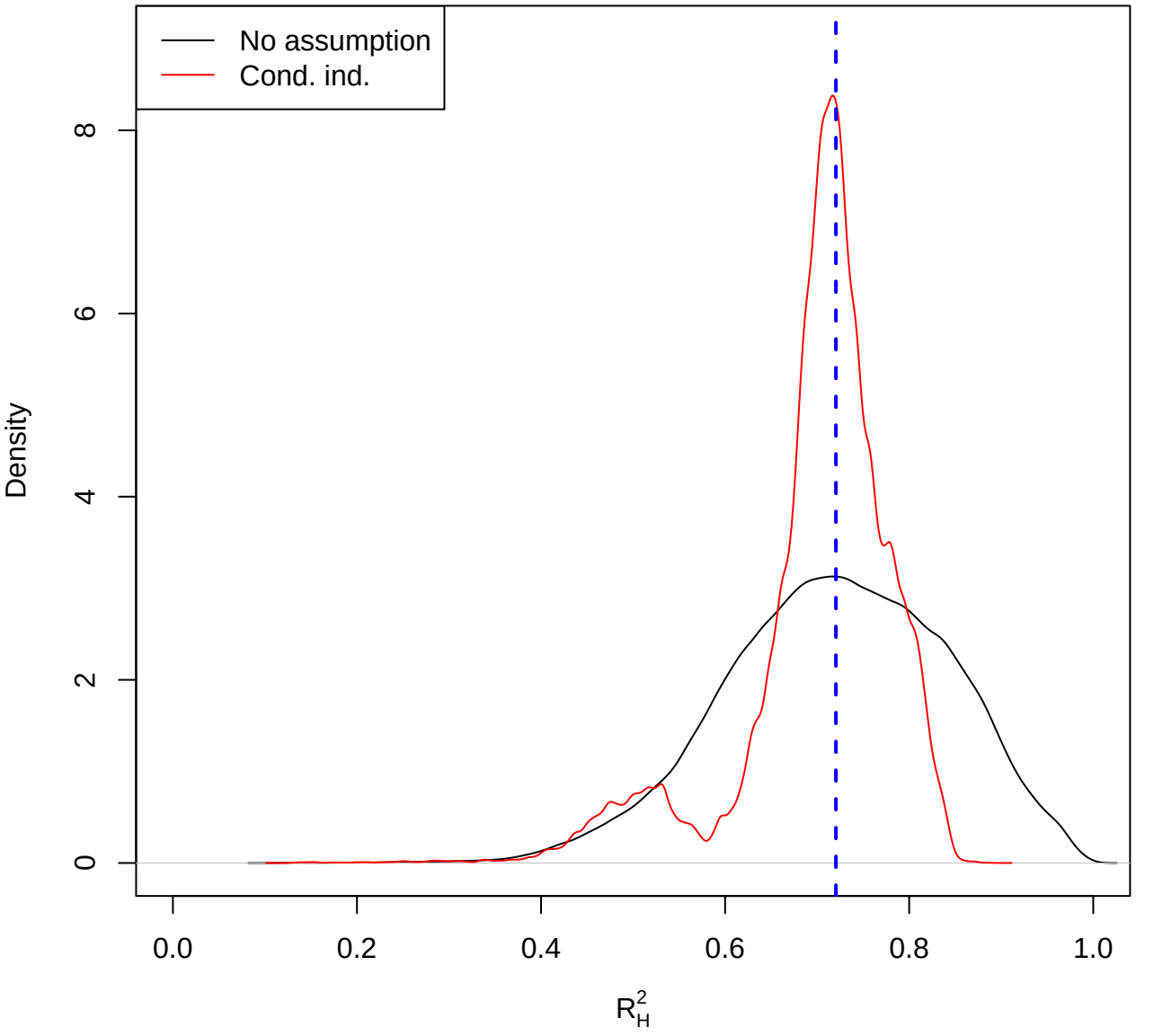

Density plot no assumption vs hvca
No assumption
HVCA
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H


Density plot no assumption vs hvca
No assumption
HVCA
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H

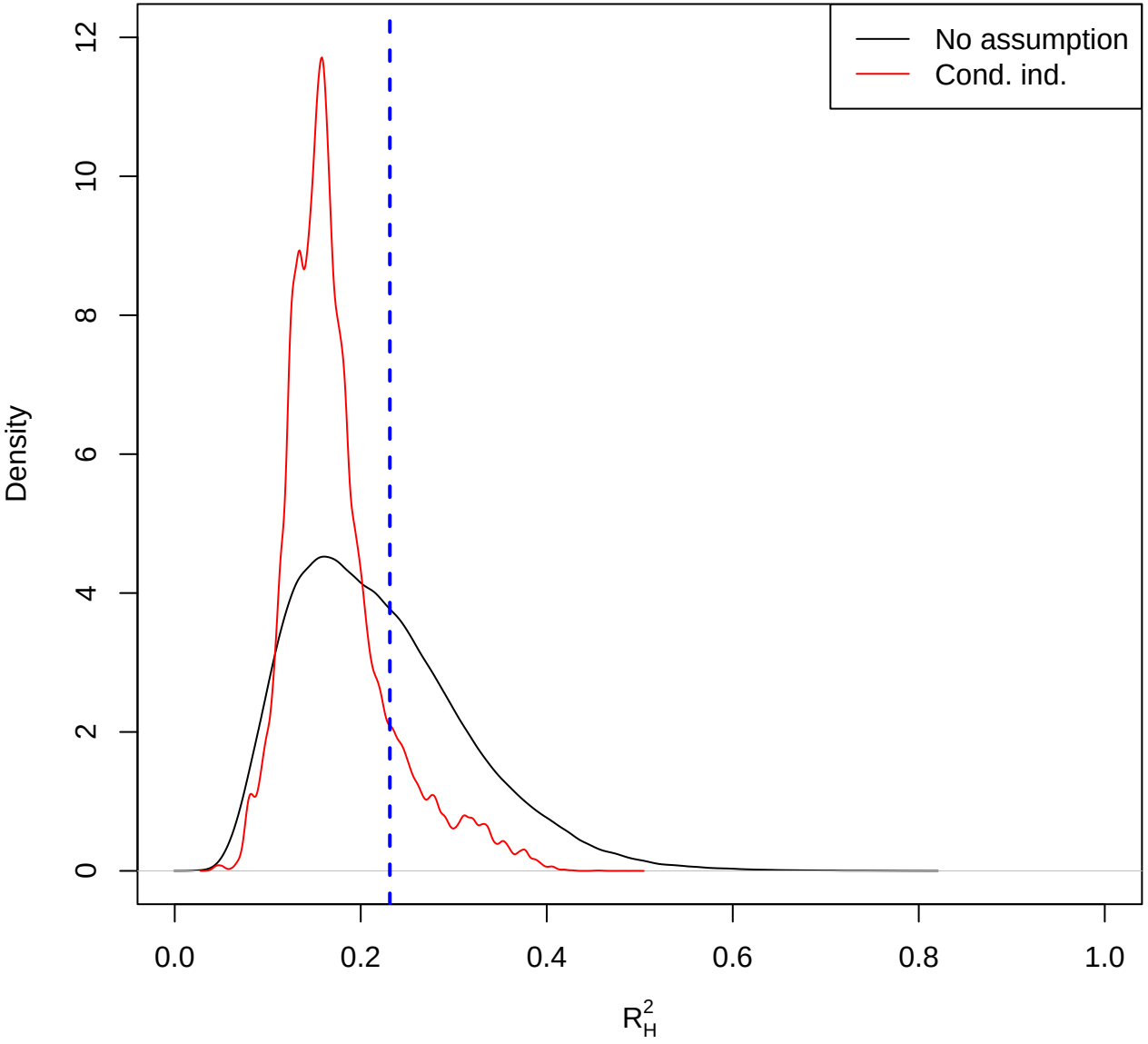
Density plot no assumption vs conditional independence



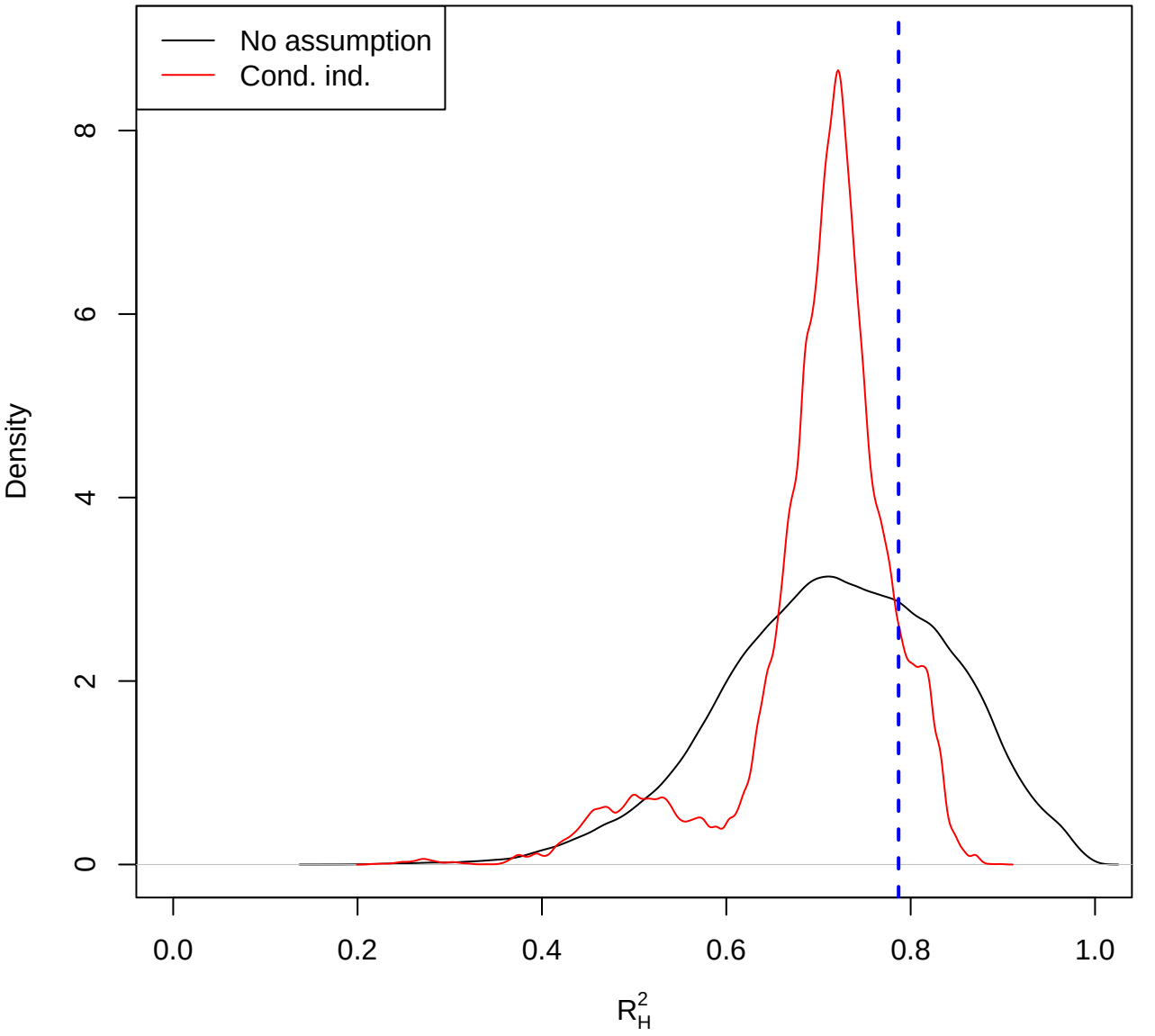
Density plot no assumption vs conditional independence



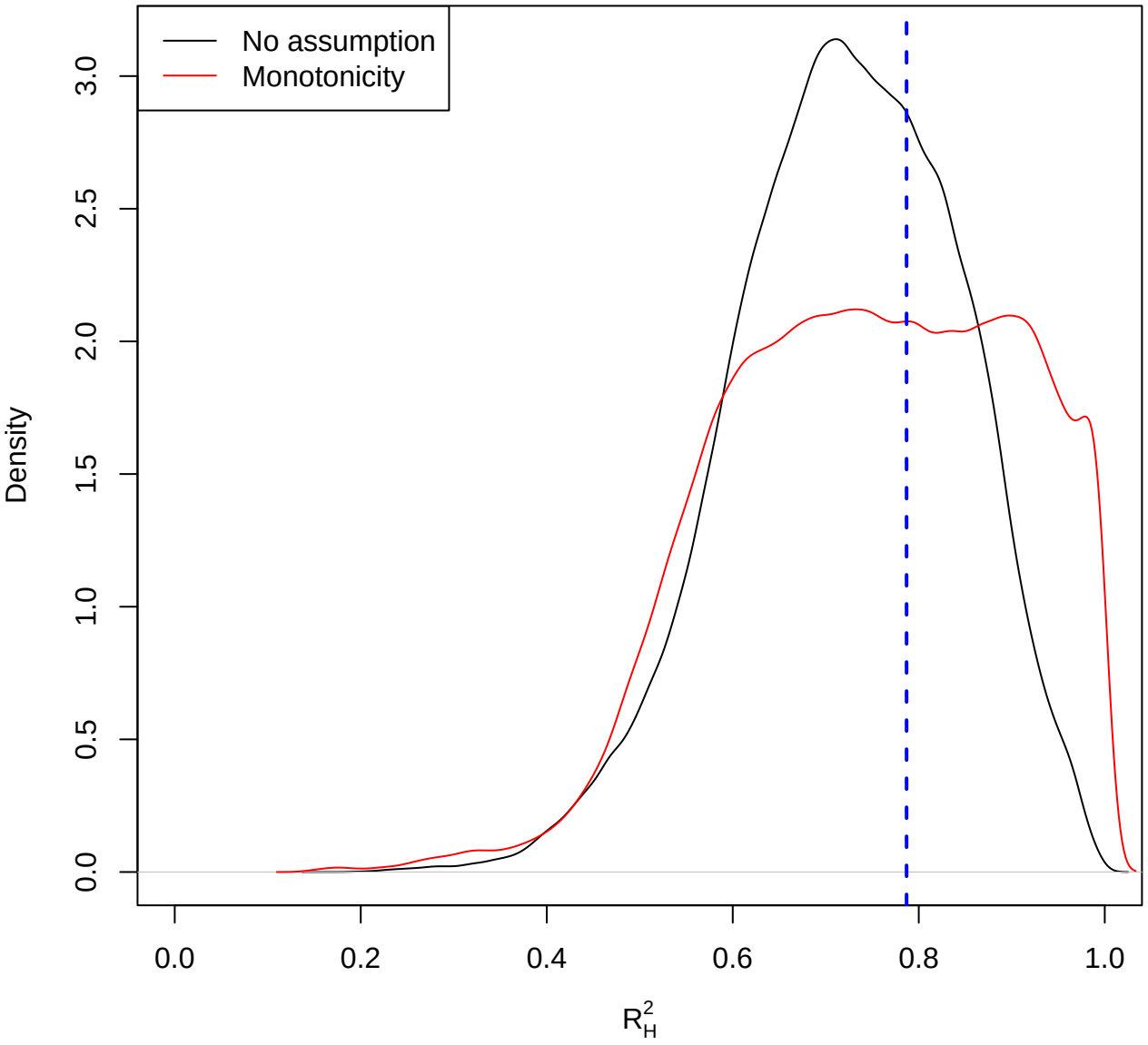
Density plot no assumption vs conditional independence

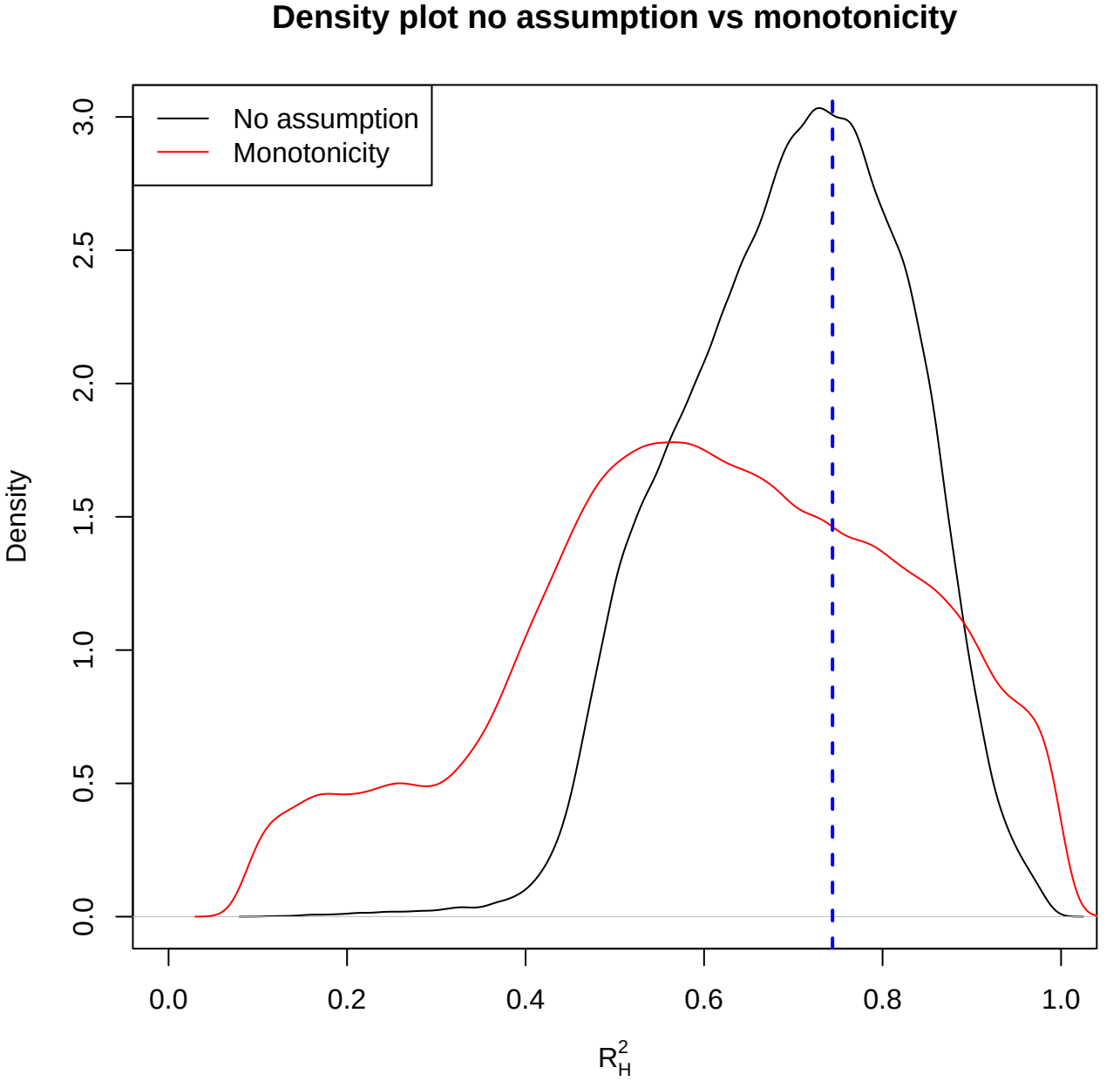


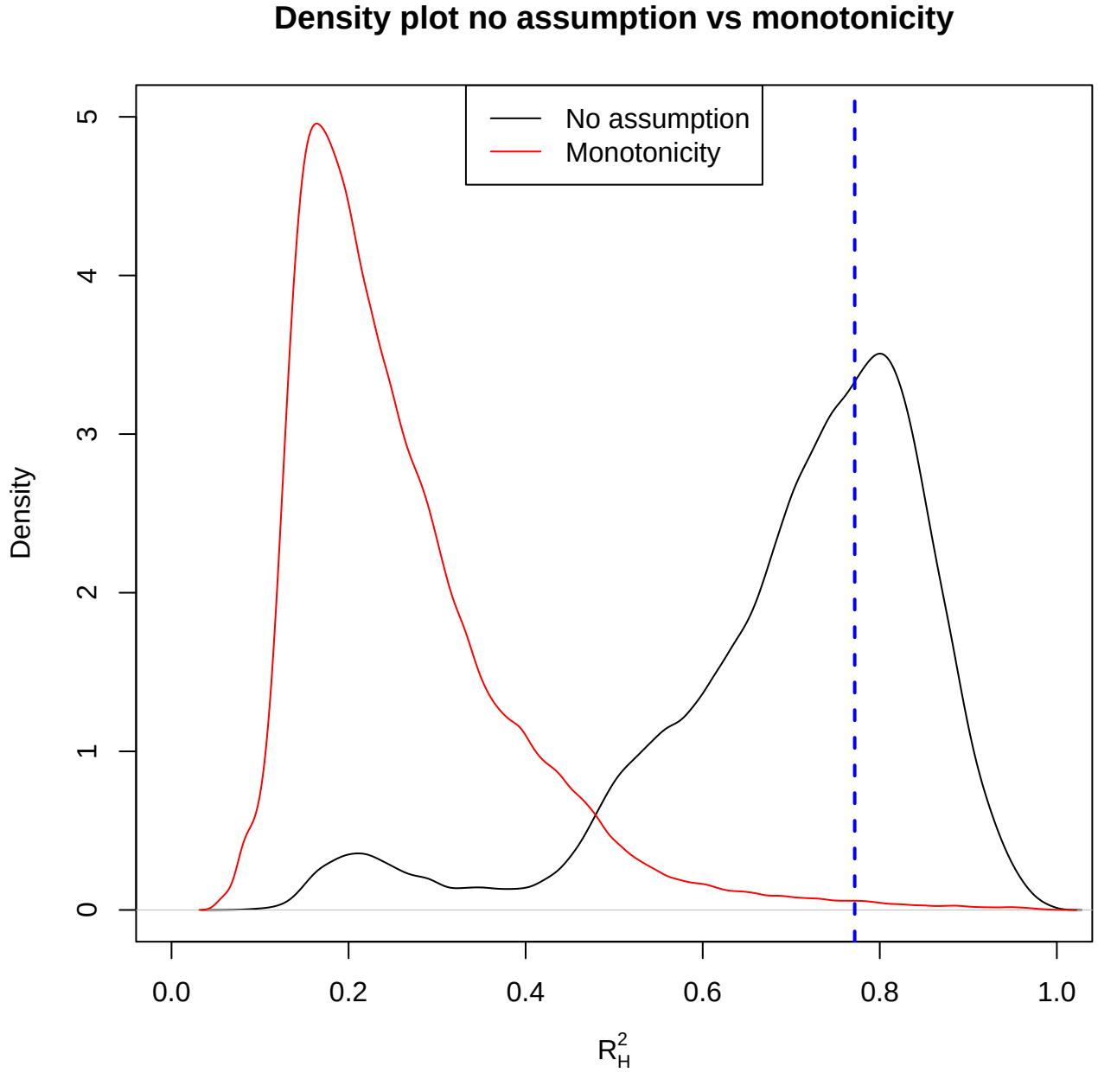
Density plot no assumption vs conditional independence

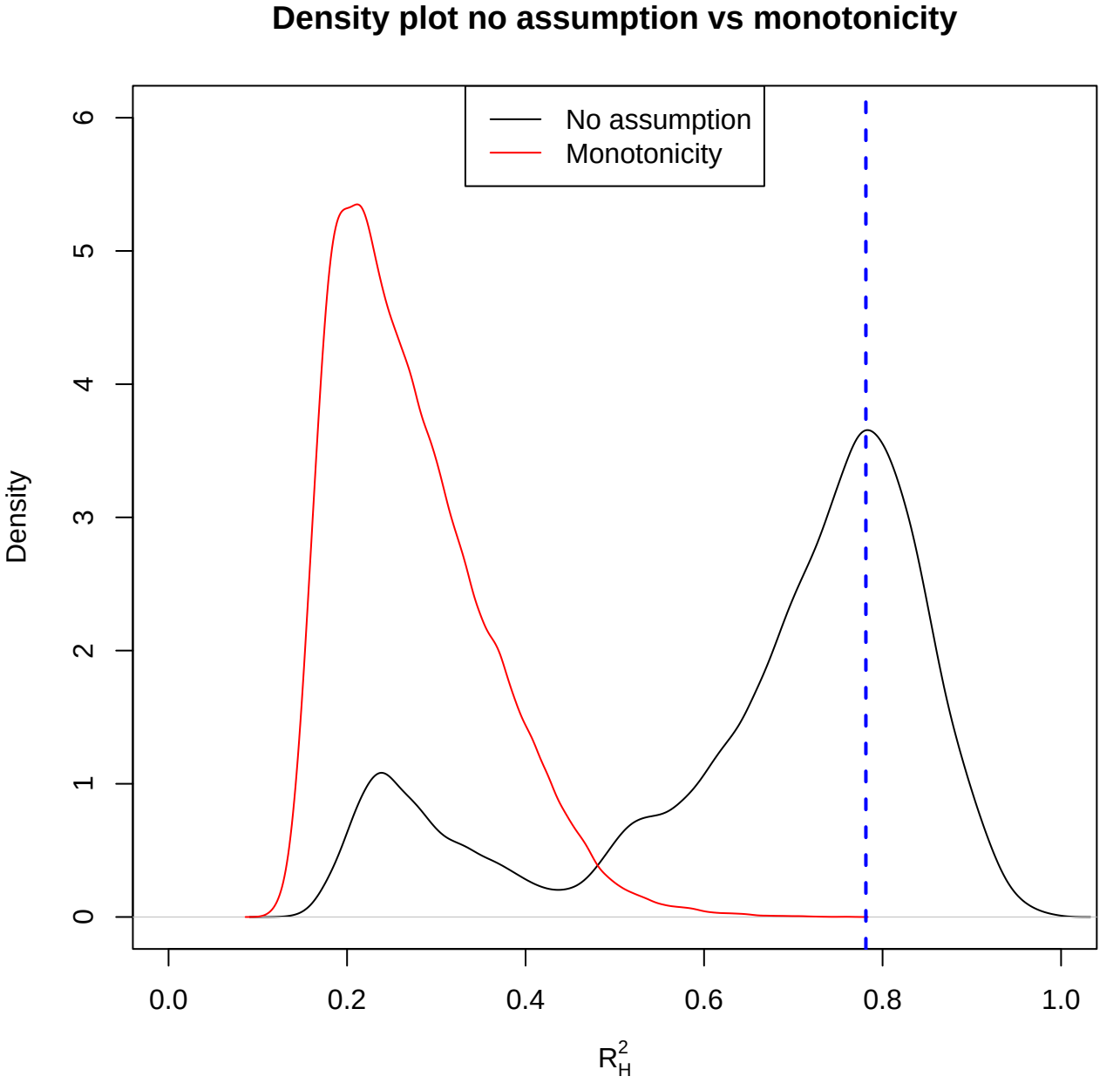


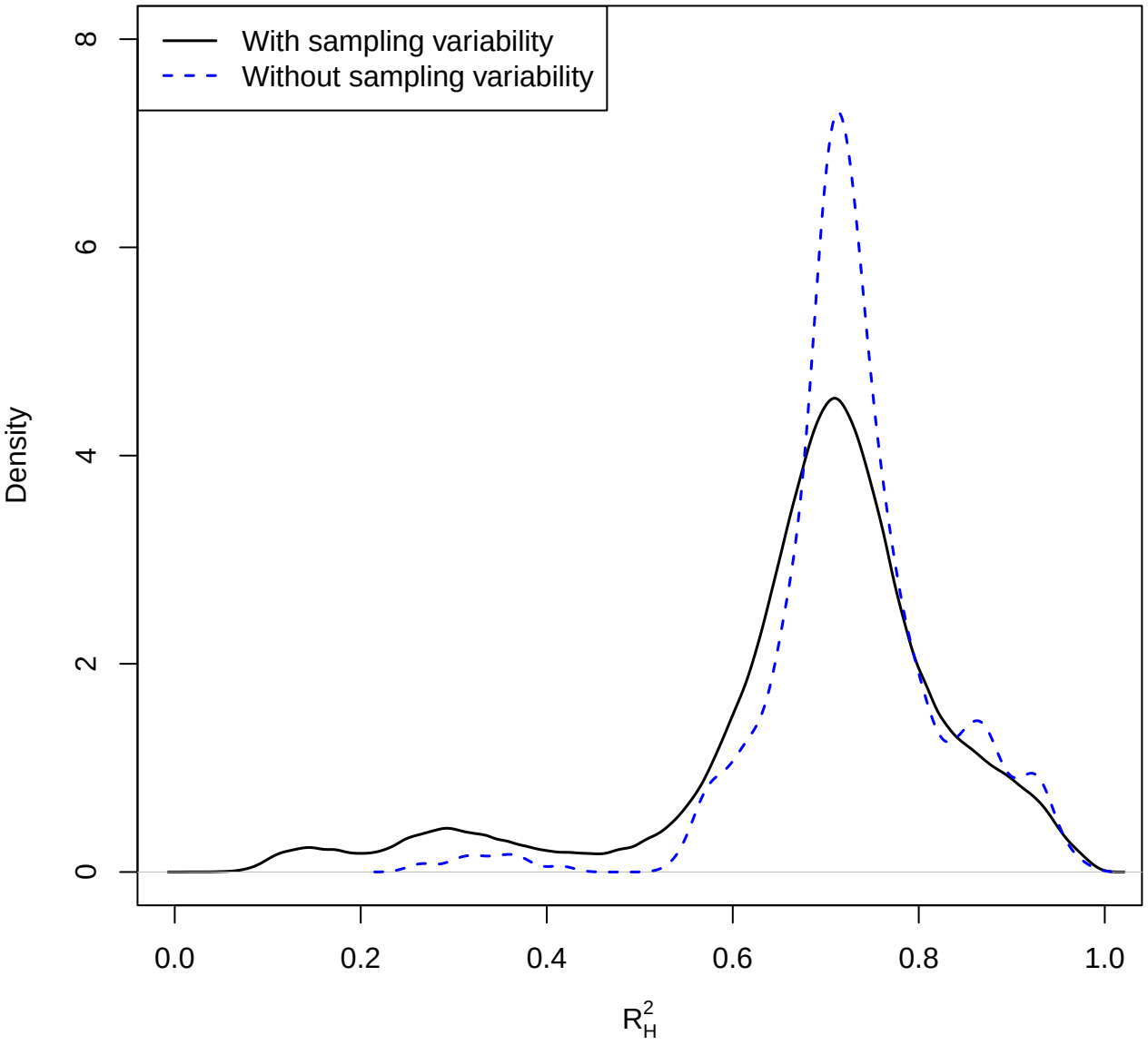
Density plot no assumption vs monotonicity




Density plot no assumption vs monotonicity
No assumption
Monotonicity
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H


Density plot no assumption vs monotonicity
No assumption
Monotonicity
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H


Density plot no assumption vs monotonicity
No assumption
Monotonicity
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H


With sampling variability
Without sampling variability
Density
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
R^2_H